TraMineR : Une boite à outils pour l'exploration et la visualisation de séquences

Gilbert Ritschard

Pôle national de recherche LIVES
Institut d'études démographiques et du parcours de vie
Université de Genève, 40, bd du Pont d'Arve, CH-1211 Genève 4, Suisse
gilbert.ritschard@unige.ch

Mots clefs : Séquences d'états, séquences d'événements, visualisation, dissimilarités, analyse basée sur les dissimilarités.

TraMineR est une librairie dévolue à l'exploration de séquences, essentiellement de séquences d'états et d'événements ordonnés chronologiquement [2]. On rencontre de telles séquences dans des domaines très divers tels que le contrôle d'appareils où l'on examine les séquences d'états de fonctionnement, en gestion où l'on s'intéresse par exemple aux successions d'achats de clients ou d'activités exercées par des employés, en analyse de l'usage du web où l'on analyse des séquences de pages visitées, et en analyse des parcours de vie où l'on étudie des séquences décrivant notamment des carrières professionnelles ou des vies familiales. Certaines des fonctionnalités proposées par TraMineR s'appliquent également à des séquences non chronologiques comme les séquences de lettres ou mots en analyse de textes, ou encore les séquences de protéines ou de nucléotides en biologie, pour lesquelles d'autres outils s'avèrent cependant mieux adaptés (voir `http://www.bioconductor.org/`).
La librairie TraMineR a été conçue à l'origine pour répondre à des questions liées à l'analyse de parcours de vie où les données comprennent typiquement quelques centaines, voire quelques milliers de séquences de longueur comprise entre 10 et 100 lorsqu'il s'agit de séquences d'états et incluant rarement plus d'une dizaine d'événements dans le cas de séquences d'événements. L'alphabet des états ou événements compte le plus souvent moins de 15 ou 20 éléments.
TraMineR offre des outils pour explorer des séquences d'états aussi bien que des séquences d'événements datés. Les séquences d'états se caractérisent par le fait que la position dans la séquence porte une information temporelle, à savoir la durée depuis le début de l'observation (par exemple le nombre de mois après la fin de la scolarité obligatoire), tandis que dans les séquences d'événements, la date de chaque événement doit être explicitement attachée à chaque événement (marié à 25 ans, premier enfant à 27 ans).

Tableau 1 – Vue transversale (gauche) versus vue longitudinale (droite)

id	t_1	t_2	t_3	$\cdots$
1	JL	JL	EM	$\cdots$
2	SC	SC	TR	$\cdots$
3	SC	SC	SC	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

id	t_1	t_2	t_3	$\cdots$
1	JL	JL	EM	$\cdots$
2	SC	SC	TR	$\cdots$
3	SC	SC	SC	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Les séquences d'états peuvent être organisées sous la forme illustrée au tableau 1 où chaque ligne correspond à un cas et chaque colonne aux unités de temps. On peut changer l'alignement en passant par exemple d'un alignement sur les dates à un alignement sur l'âge (durée du processus). Les outils proposés pour les séquences d'états permettent en particulier de

– rendre compte de l'évolution des distributions transversales (chronogrammes, entropies transversales, ...) ;
– visualiser l'ensemble des séquences individuelles et de calculer des caractéristiques (nombre de changements d'états, durées moyennes dans les états ou complexité de la séquence par exemple) ;
– calculer la dissimilarité entre séquences selon plusieurs métriques ;
– visualiser des groupes et donc en particulier les clusters qui peuvent être déduits des dissimilarités.

La librairie offre également plusieurs outils d'analyse originaux fondés sur les dissimilarités :
– calculer et analyser la dispersion des séquences [6] ;
– identifier visualiser des séquences représentatives (medoid, avec plus forte densité, ...) [3] ;
– générer des arbres de régression de séquences [6].

Les séquences d'événements se distinguent des séquences d'états par l'absence d'alignement sur une date ou un âge et la possibilité d'avoir des événements simultanés. Les outils spécifiques offerts sont [4] :
– visualisation sous forme de 'parallel coordinate plot' [1] ;
– extraction de sous-séquences fréquentes sous diverses contraintes de temps, de contenu et de selon diverses méthodes de comptage ;
– identification des sous-séquences les plus discriminantes entre groupes ;
– calcul de dissimilarités entre séquences d'événements.

La librairie inclut également plusieurs fonctions utilitaires notamment pour convertir entre diverses possibilités d'organisation des données et en particulier pour aider à convertir entre séquences d'états et séquences d'événements datés [5].

La présentation portera sur la genèse de la librairie et sa philosophie axée sur des objets séquences d'états et séquences d'événements qui incluent un maximum d'information comme l'alphabet, les étiquettes courtes et longues, la palette de couleur pour les visualisation et les pondérations pour n'en citer que quelques uns. Nous évoquerons également l'attention accordée à la documentation et au support offert aux utilisateurs.

Références

[1] Bürgin, R. and G. Ritschard (2012). Categorical parallel coordinate plot. In *LaCOSA Lausanne Conference On Sequence Analysis, University of Lausanne, June 6th-8th 2012*, Lausanne. Poster.

[2] Gabadinho, A., G. Ritschard, N. S. Müller, and M. Studer (2011a). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software* **40**(4), 1–37.

[3] Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2011b). Extracting and rendering representative sequences. In A. Fred, J. L. G. Dietz, K. Liu, and J. Filipe (Eds.), *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Volume 128 of *Communications in Computer and Information Science (CCIS)*, pp. 94–106. Springer-Verlag.

[4] Ritschard, G., R. Bürgin, and M. Studer (2013). Exploratory mining of life event histories. In J. J. McArdle and G. Ritschard (Eds.), *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, Quantitative Methodology. New York : Routledge. (in press)

[5] Ritschard, G., A. Gabadinho, M. Studer, and N. S. Müller (2009). Converting between various sequence representations. In Z. Ras and A. Dardzinska (Eds.), *Advances in Data Management*, Volume 223 of *Studies in Computational Intelligence*, pp. 155–175. Berlin : Springer-Verlag.

[6] Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research* **40**(3), 471–510.