Exploring Sequential Data

Gilbert Ritschard

NCCR LIVES and Institute for Demographic and Life Course Studies,
University of Geneva, CH-1211 Geneva 4, Switzerland
gilbert.ritschard@unige.ch

Abstract. The tutorial is devoted to categorical sequence data describing for instance the successive buys of customers, working states of devices, visited web pages, or professional careers. Addressed topics include the rendering of state and event sequences, longitudinal characteristics of sequences, measuring pairwise dissimilarities and dissimilarity-based analysis of sequence data such as clustering, representative sequences, and regression trees. The tutorial also provides a short introduction to the practice of sequence analysis with the TraMineR R-package.

1 Types of Categorical Sequence Data

Categorical sequence data are present in very different fields. There are sequences with a chronological order such as in device control where we have sequences of successive activity events [10], management where we have sequences of successive goods bought by customers [3] or sequences of types of activity carried out by employees, in web usage analysis where we have sequences of visited pages [9], and in life course studies where we have sequences describing work careers or family life trajectories [2, 15]. In other domains sequences do not have a time dimension. This is for example the case of sequences of proteins or nucleotides, or of sequences of letters and words in texts.

The kind of knowledge we are interested in discovering from the sequential data varies across disciplines. In biology the aim is to find out repeating patterns in a same sequence or known patterns in given sequences. In text analysis, the aim could be to find out patterns that characterize an author, or patterns common to texts dealing with a specific subject. The tutorial does not cover these latter aspects, but focuses on methods primarily intended for sequences with a time dimension. We consider methods for rendering and exploring a series of hundreds or even thousands sequences of length more or less between 10 and 100, and an alphabet of symbols of limited size, say less than 20.

A categorical sequence is a ordered list of symbols chosen from a given alphabet. For example, we will consider data describing the transition from school to work of Irish students [11] which indicate in which of six states (EM = ‘employment’, FE = ‘further education’, HE = ‘higher education’, JL = ‘joblessness’, SC = ‘school’, TR = ‘training’) the students are during each of the 72 months following the end of compulsory school. For that example, the alphabet is of size six and the sequences are of length 72. There are 762 observed sequences.

J.-G. Ganascia, P. Lenca, and J.-M. Petit (Eds.): DS 2012, LNAI 7569, pp. 3–6, 2012.

Table 1. Transversal view (left) versus longitudinal view (right)

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	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Such state sequences typically correspond to panel data and can be organized in tabular form (Table 1) with each row corresponding to a case (a sequence) and each column to a position in the sequence. For state sequences, the position conveys time information, namely the duration from the start of the sequence. Instead of states, we can consider the events that provoke the change of state. Changing from school to employment, for example, supposes that we end schooling and start a job the same months. Event sequences differ from state sequences in two aspects: (1) the position of the events in the sequence do not convey time information. Explicit time stamps are needed if we want account for time. (2) Multiple events can occur simultaneously, while states are mutually exclusive. The tutorial addresses methods for state sequences only. The presentation will mainly consist in commented examples. We also shortly demonstrate how easily such analyses can be run in R with the TraMineR toolbox [7, 14]. Nevertheless, the TraMineR package also provides tools for event sequences. See for instance [13] to get an idea of the kind of results you may derive from event sequences.

We start by discussing descriptive statistics of state sequence data and then turn to more advanced dissimilarity-based analysis of sequential data.

2 Describing State Sequences

State sequences can be analyzed from two complementary standpoints. We can look at the transversal distributions (Table 1 left). sequences of transversal summaries (modal states, transversal entropies) give an aggregated view of the time evolution of the set of sequences. Alternatively, we can look at the longitudinal characteristics of each sequence (Table 1 left). Alongside with plots for rendering those different aspects, the tutorial shortly discusses useful characteristics of transversal distributions and the longitudinal characteristics of individual sequences (number of transitions, longitudinal entropy, index of complexity, turbulence).

3 Dissimilarity-Based Analysis

The success of sequence analysis in the social sciences is largely attributable to Abbott [1] who introduced the so called 'Optimal Matching' (OM) analysis to sociologists and historians. OM analysis consist in computing pairwise dissimilarities between sequences by means of an edit distance and then running a clustering analysis from the obtained dissimilarities. Hamming, distance based on the longest common subsequence, on the number of common subsequences, ... there are alternatives for computing such dissimilarities. Whichever we use, it open access not only to clustering but to a whole range of dissimilarity-based analysis: multidimensional-scaling, identifying sequences with the densest neighborhoods [8], discrepancy analysis and regression trees for sequence data [14].

4 Conclusion

This tutorial gives a very short introduction to sequence analysis as it is practiced for life course analysis. All the methods addressed are available in TraMineR. The current users of TraMineR come from a great variety of disciplines and the methods have been used for example in studies of invertebrate movements [16], of disease management [6], in political science [4, 5], for web usage analysis [9] as well as in analysis of impact of feedback on mobile interaction with maps [12]. The TraMineR package is available from the CRAN (`http://cran.r-project.org/web/packages/TraMineR`). For more details about the package, see `http://mephisto.unige.ch/traminer`.

References

[1] Abbott, A., Forrest, J.: Optimal matching methods for historical sequences. Journal of Interdisciplinary History 16, 471–494 (1986)

[2] Abbott, A., Tsay, A.: Sequence analysis and optimal matching methods in sociology, Review and prospect. Sociological Methods and Research 29, 3–33 (2000); (With discussion, pp. 34–76)

[3] Agrawal, R., Srikant, R.: Mining sequential patterns. In: Yu, P.S., Chen, A.L.P. (eds.) Proceedings of the International Conference on Data Engeneering (ICDE), Taipei, Taiwan, pp. 487–499. IEEE Computer Society (1995)

[4] Buton, F., Lemercier, C., Mariot, N.: The household effect on electoral participation. A contextual analysis of voter signatures from a french polling station (1982-2007). Electoral Studies 31, 434–447 (2012); Special Symposium: Generational Differences in Electoral Behaviour

[5] Casper, G., Wilson, M.: Bargaining within crises. In: American Political Science Association Meetings, Seattle, WA, September 1-4 (2011)

[6] Donnachie, E., Hofman, F., Keller, M., Mutschler, R., Wolf, R.: Qualitätsbericht 2010: Disease management programme in bayern. Bericht, Gemeinsame Einrichtung DMP Bayern, Bayern (D) (2011)

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

[8] Gabadinho, A., Ritschard, G., Studer, M., Müller, N.S.: Extracting and rendering representative sequences. In: Fred, A., Dietz, J.L.G., Liu, K., Filipe, J. (eds.) IC3K 2009. CCIS, vol. 128, pp. 94–106. Springer, Heidelberg (2011)

[9] Jiang, Q., Tan, C.H., Phang, C.W., Wei, K.K.: Using sequence analysis to classify web usage patterns across websites. In: Hawaii International Conference on System Sciences, pp. 3600–3609 (2012)

[10] Mannila, H., Toivonen, H., Verkamo, A.I.: Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery 1, 259–289 (1997)

[11] McVicar, D., Anyadike-Danes, M.: Predicting successful and unsuccessful transitions from school to work using sequence methods. Journal of the Royal Statistical Society A 165, 317–334 (2002)

[12] Reilly, D.F., Inkpen, K.M., Watters, C.R.: Getting the picture: Examining how feedback and layout impact mobile device interaction with maps on physical media. In: IEEE International Symposium on Wearable Computers, pp. 55–62 (2009)

[13] Ritschard, G., Bürgin, R., Studer, M.: Exploratory mining of life event histories. In: McArdle, J.J., Ritschard, G. (eds.) Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences. Quantitative Methodology. Routledge, New York (2012)

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

[15] Widmer, E., Ritschard, G.: The de-standardization of the life course: Are men and women equal? Advances in Life Course Research 14, 28–39 (2009)

[16] Zou, S., Liedo, P., Altamirano-Robles, L., Cruz-Enriquez, J., Morice, A., Ingram, D.K., Kaub, K., Papadopoulos, N., Carey, J.R.: Recording lifetime behavior and movement in an invertebrate model. PLoS ONE 6, e18151 (2011)