Robust timestamps and privacy preservation for inference from longitudinal data

Anna Palczewska (A.Palczewska@leeds.ac.uk)
joint work with Georgios Aivaliotis and Jan Palczewski
University of Leeds

The widespread collection of data in social, health and consumer contexts has contributed to the availability of complex temporal datasets. Data instances collected in these datasets are characterised by a variable number of irregularly spaced timestamped events that include information about continuous activities and instantaneous events (e.g. Electronic Health Records). The increase in the number of complex temporal datasets has prompted the development of methods that extend applicability of classical statistical, machine learning and data mining methods to those datasets [1, 3, 4]. These methods involve identification of relevant temporal patterns of events. Patterns are used to encode original variable-length data instances as fixed-length binary vectors representing presence or absence of chosen patterns therefore enabling application of existing classification and prediction tools.

This paper makes two contributions. Firstly, we design a robust approach for analysis of longitudinal data that can account for noise/errors in the recorded timestamps. Such errors are common in information systems maintained manually such as health records or social care systems. Our contribution is both theoretical (formulation of robust temportal patterns) and algorithmic (efficient methods for identification of such robust patterns which go beyond and are significantly harder than classical problems of sequential pattern mining).

Our second contribution concerns privacy preservation/statistical disclosure problems in the context of longitudinal data. As such data contain rich temporal information, they are at risk of being used in person re-identification [5]. A typical approach of using aggregate statistics is often insufficient for complex datasets as experience in [2] shows. Instead, we borrow from a classical approach in statistical disclosure and perturb timestamps with a random noise. Analysis with such perturbed data is possible through our robust timestamps approach. We will report on the performance of this method in predictive modelling; in particular, we will discuss the loss of predictive power and the sensitivity to specification of the noise.

As a test ground for our methods we will use a risk stratification problem in Adult Social Care. The dataset is provided by Leeds City Council and contains for each client: timestamped referrals, assessments, reviews and services provided as well as static health and socio-economic descriptors. The aim is to identify clients that are at the highest risk of moving into expensive care (such as nursing or residential care) in order to provide additional services to extend their independence.

This research is funded by EPSRC through grant no EP/N013980/1: "QuantiCode: Intelligent infrastructure for quantitative, coded longitudinal data".

References

[1] I. Batal et al. *An efficient pattern mining approach for event detection in multivariate temporal data*, Knowledge and Information Systems 46, pp. 115-150, 2016

[2] M. Bardsley et al. *Predicting social care costs: a feasibility study*, Nuffield Trust, 2011

[3] Yi-Cheng Chen et al., *A novel algorithm for mining closed temporal patterns from interval-based data*,Knowledge and Information Systems 46, pp. 151-183, 2016

[4] R. Moskovitch, Y. Shahar, *Classification of multivariate time series via temporal abstraction and time intervals mining*,Knowledge and Information Systems 45, pp. 35-74, 2015

[5] J. Domingo-Ferrer, R. Trujillo, *Anonymization of trajectory data*. Joint UNECE—Eurostat work session on statistical data confidentiality, Tarragona, Spain, 26-28 October 2011