Event History Analysis on Data Streams: An Application to Earthquake Occurrence

Ammar Shaker and Eyke Hüllermeier

Department of Mathematics and Computer Science
Marburg University, Germany
{shaker,eyke}@mathematik.uni-marburg.de

1 Background: Earthquake Data

Earthquakes are natural disasters with an effect proportional to their causalities and caused destruction. Although the media report about only a few earthquakes in different regions of the world every year, earthquakes actually occur much more frequently, with varying impacts on their surrounding regions. One of the trusted sources of information about earthquakes is the USGS (United States Geological Survey) and its section NEIC (National Earthquake Information Center), whose missions are to quickly discover the most recent destructive earthquakes in terms of location and magnitude, and to broadcast this information to international agencies and scientists.[1] The USGS offers an online catalog, the ANSS Comprehensive Catalog (ComCat), which stores information about earthquake source parameters (e.g., hypocenters, magnitudes, phase picks and amplitudes) as well as other summaries and moments. This data is produced from a large network of seismic stations scattered around the globe.

The seismic signals of an earthquake reach few seismic stations, which manage to locate its hypocenter by analyzing the seismic P-wave, one of the different seismic waves created by an earthquake. The P-wave is the fastest wave, spreading with a speed of about $5-8$ km/s. From the list of the arrival times of P-waves at different stations, an estimate of the hypocenter is produced. Another type of seismic wave is the S-wave, which spreads slower than the P-waves in about $3-5$ km/s. Despite being slow, these waves hold most of the seismic energy generated by the earthquake, which can be used to estimate its magnitude. Different updates on the location and the magnitude of an earthquake are performed as more information is gathered. These updates are done in the next hours and days after the earthquake. Further updates on the estimated magnitude are possible after more sophisticated analysis.

2 Challenges

From a stream analysis point of view, earthquakes provide a potentially interesting source of data, especially since earthquake data is produced continuously

[1] http://www.usgs.gov/, http://earthquake.usgs.gov/regional/neic/

in the course of time. More importantly, this type of data exhibits a number of characteristics that make is specifically challenging:

1. Event data: The data produced by earthquake monitoring essentially corresponds to *event data*, i.e., information about the time points at which a specific type of event occurs. This type of temporal data is well known in statistics but has received little (or actually no) attention in data mining and machine learning (on streams) so far.
2. Spatio-temporal analysis: The data does not only have a temporal but also a spatial dimension, namely the location of an earthquake. Considering both dimensions simultaneously leads to spatio-temporal data analysis, which has been studied extensively in the static setting but much less in a dynamic or online setting. Moreover, the effect of the spatial dimension is highly nonlinear and hence calls for an appropriate modeling of its influence.
3. Delayed observations: Information about earthquakes often arrives with a delay of non-constant length. These delays could range from hours to days, as they are provided by a wider contributing network of seismic stations.

We elaborate on each of these challenges in Sections 3, 4 and 5, respectively.

3 Event History Analysis on Data Streams

Event history analysis is an established statistical method for the study of temporal "events" or, more specifically, questions regarding the temporal distribution of the occurrence of events and their dependence on covariates of the data sources. In [Shaker and Hüllermeier, 2013], we made a first step toward analyzing this type of data in the context of data streams. To this end, we develop an incremental, adaptive version of *event history analysis* (EHA), which is a standard statistical method for event analysis. The basic mathematical tool in EHA is the *hazard function*, which models the "propensity" of the occurrence of an event (marginal probability of an event conditional to no event so far) as a function of time.

To the best of our knowledge, EHA has not received much attention in the stream setting so far, which is arguably surprising for several reasons. Most notably, the temporal nature of event data naturally fits the data stream model, and indeed, "event data" is naturally produced by many data sources, including but not limited to earthquake data.

To make event history analysis applicable in the setting of data streams, we develop an adaptive (online) variant of a model that is closely related to the well-known proportional hazard model proposed by Cox [Cox and Oakes, 1984]. In this model, the hazard rate may depend on one or more covariates associated with a statistical entity. More specifically, in the proportional hazard model, the effect of an increase of a covariate by one unit is multiplicative with respect to the hazard rate. We estimate the influence of the covariates by adopting the sliding window approach, assuming that the hazard rate is constant on every window. To this end, an online version of a maximum likelihood estimation procedure has been developed.

4 Event History Analysis for Earthquake Data

We applied our streaming version of EHA to the analysis of earthquake data. The earthquakes were collected in the time period between the 1st Jan 2000 and the end of 27th Mar 2012. In total, we collected 319,884 earthquakes around the entire globe. While the temporal aspect is naturally captured by the hazard rate model, the spatial aspect is incorporated through the use of spatial information as covariates of the data streams. In other words, the vector of covariates is describing the spatial location of a data source.

In our setting, we assume to observe event sequences for a fixed set of sources, each of which corresponds to a data stream. In order to define these sources, we discretize the globe both in terms of longitude and latitude, and associate one source with each intersection point. Moreover, in order to account for possibly nonlinear dependencies between spatial coordinates and risk of earthquake, we define the features of the data sources in terms of a fuzzy partition, that is, a partition defined in terms of fuzzy sets as shown in Figure 1(c). In contrast to a standard partition defined in terms of intervals, this allows for a smooth transition between spatial regions. A two-dimensional (fuzzy) discretization of the globe is defined in terms of the Cartesian product of these two one-dimensional discretizations, using the minimum operator for fuzzy set intersection.

Deriving time-dependent estimates of the model parameters on time windows of 180 days shifting every 30 days, several interesting observations could be made for data from the last decade. For example, as can be seen in Figure 1(a), the occurrence of Tohoku's earthquake in March 2011 comes with a significant increase in the coefficients of the fuzzy sets covering that area. The coefficient of the green line increases by a factor of 4 till few hours before the earthquake, indicating an increased hazard rate for the area where the earthquake has occured.

5 Dealing with Time Delays

As already mentioned, information about earthquakes may arrive with a certain time delay, and information about past events might be updated in the course of time when more accurate information is available.

To get an idea of the relevance of this problem, we took different snapshots of all available earthquakes that occurred during the time interval from $A =$ 1-Jan-2000 till $B =$ 21-May-2013. Repeating this operation every half an hour for about one month, we generate snapshots of earthquakes on the intervals $[A, B + i \times 30 \text{ min}]$, $i \in \{0, 1, \ldots, 1440\}$. We can then compare pairs of snapshots s_1 and s_2 taken for periods $[a, b]$ and $[a, b+\Delta]$ on the shared time interval $[a-\delta, a]$. This comparison reveals how many events occurring in $[a - \delta, a]$ are contained in snapshot s_2 while being absent from s_1. Thus, it allows for discovering the number of events coming with a delay. Figure 1(c) shows different combinations of δ and Δ, with δ on the horizontal axes and assigning different colors to the different values of Δ. As an example, we can see that an average of 193.03 earthquakes are observed with a latency bounded by 6 days. Needless to say, by

simply ignoring delays of that extent, the results of EHA on data streams might be strongly biased. The question of how to handle time delays in a proper way is part of ongoing work, but a convincing solution has not yet been found. In fact, learning from delayed feedback has not received much attention in the data stream community so far, despite existing work in the related setting of online learning (e.g. [Joulani et al., 2013]).

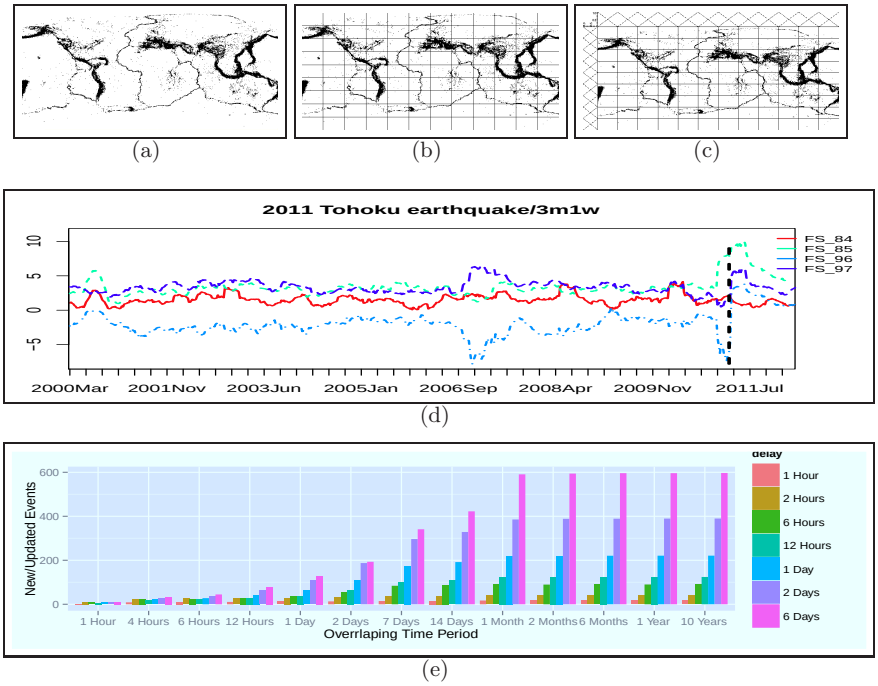

Fig. 1. The collected earthquakes plotted by their geographic coordinates. (a) earthquakes only; (b) with coordinates lines (lat,lng); (c) fuzzy partitions on the two coordinates. (d) The coefficients of features for the region of the 2011 significant earthquake; (e) The average number of delayed events categorised by their delay.

References

[Cox and Oakes, 1984] Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

[Joulani et al., 2013] Joulani, P., György, A., and Szepesvári, C. (2013). Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, USA*.

[Shaker and Hüllermeier, 2013] Shaker, A. and Hüllermeier, E. (2013). Event history analysis on data streams. *International Journal of Applied Mathematics and Computer Science*, under revision.