Towards Understanding and Predicting Distributional Changes

Georg Krempl* Dominik Lang Vera Hofer Katarzyna Bijak
Utrecht University Magdeburg University Graz University University of
Dept. Information Faculty of Dept. of Statistics & Southampton
Princetonplein 5 Universitätspl. 2 Universitätsst. 15/E3 Highfield
3584 CC Utrecht 39106 Magdeburg 8010 Graz SO17 1BJ
The Netherlands Germany Austria United Kingdom
* g.m.krempl@uu.nl

Abstract

An important foundation for credit scoring, as for statistics and machine learning in general, is making inference about future sample values from analysing historical training samples of data. However, distributions are often not stationary but change over time, for example due to changing economic conditions [Kelly et al., 1999, Zliobaite et al., 2016]. In recent years, techniques have been developed to update prediction models on streams of data, either continuously or upon detection of change [Gama et al., 2014]. While most techniques rely on new labelled data, a few approaches have been proposed that aim to exploit patterns in the distributional change to update prediction models [Plasse and Adams, 2016, Tasche, 2014, Dyer et al., 2014], and [Hofer and Krempl, 2013], [Krempl, 2011]. However, an important step towards reliable adaptive prediction techniques is to better understand and model changes in distributions over time.

In this work, we present the Temporal Density Extrapolation approach for this task. This novel method models and predicts gradual monotonous changes in a distribution. It is based on the expansion of basis functions. The weights of these basis functions are modelled as functions of compositional data over time by using an isometric log-ratio transformation. To obtain extrapolated density estimates, the weights are extrapolated to the requested time point, and the density is queried from the basis functions with back-transformed weights. This approach is designed for broad applicability by neither being restricted to a specific parametric distribution, nor relying on cluster structure in the data. It requires only two additional extrapolation-specific parameters, for which reasonable defaults exist.
Results of an experimental evaluation on various data streams, which comprises synthetic and real-world data, including credit scoring data, show that the model manages to capture monotonous drift patterns accurately and better than existing methods. Thereby, it requires not more than 1.5 times the run time of a corresponding static density estimation approach. This makes it an interesting candidate for analysing and predicting continuous distributional change patterns in time-evolving streaming data.

References


