MixedTrails: Bayesian Hypothesis Comparison Heterogeneous Sequential Data

Martin Becker[1], Philipp Singer[2] Florian Lemmerich[2],
Andreas Hotho[1,3], and Markus Strohmaier[2,4]

[1] University of Würzburg, Germany
{becker,hotho}@informatik.uni-wuerzburg.de
[2] GESIS, Cologne, Germany
{philipp.singer,florian.lemmerich,markus.strohmaier}@gesis.org
[3] L3S Research Center, Hannover, Germany
{philipp.singer,markus.strohmaier}@gesis.org
[4] University of Koblenz-Landau, Mainz, Germany

Sequential traces of user data are frequently observed online and offline, e.g., as sequences of visited websites or as sequences of locations captured by GPS. However, understanding factors explaining the production of sequence data is a challenging task, especially since the data generation is often not homogeneous. For example, navigation behavior might change in different phases of browsing a website, or movement behavior may vary between groups of users. In this work, we tackle this task and propose *MixedTrails* [1], a Bayesian approach for comparing the plausibility of hypotheses regarding the generative processes of heterogeneous sequence data. Each hypothesis is derived from existing literature, theory or intuition and represents a belief about transition probabilities between a set of states that can vary between groups of observed transitions. For example, when trying to understand human movement in a city, a hypothesis assuming tourists to be more likely to move towards points of interests than locals, can be shown to be more plausible with observed data than a hypothesis assuming the opposite. Our approach incorporates these beliefs as Bayesian priors in a generative mixed transition Markov chain model, and compares their plausibility utilizing Bayes factors. We discuss analytical and approximate inference for calculating the marginal likelihoods for Bayes factors, give guidance on interpreting the results, and illustrate our approach with several experiments on synthetic and empirical data from Wikipedia and Flickr. Thus, this work enables a novel kind of analysis for studying sequential data in many application areas.

References

1. Becker, M., Lemmerich, F., Singer, P., Strohmaier, M., Hotho, A.: Mixedtrails: Bayesian hypothesis comparison on heterogeneous sequential data. Data Mining and Knowledge Discovery (Jul 2017)

In: M. Leyer (Ed.): Proceedings of the LWDA 2017 Workshops: KDML, FGWM, IR, and FGDB.
Rostock, Germany, 11.-13. September 2017, published at http://ceur-ws.org