Comparing hypotheses about sequential data: A Bayesian approach and its applications

Florian Lemmerich[1], Philipp Singer, Martin Becker[2], Lisette Espin-Noboa[1], Dimitar Dimitrov[1], Denis Helic[3], Andreas Hotho[2], and Markus Strohmaier[1,4]

[1] GESIS - Leibniz Institute for the Social Sciences
firstname.lastname@gesis.org
[2] University of Würzburg
lastname@informatik.uni-wuerzburg.de
[3] Graz University of Technology
dhelic@tugraz.at
[4] RWTH Aachen
markus.strohmaier@humtec.rwth-aachen.de

Abstract. Sequential data can be found in many settings, e.g., as sequences of visited websites or as location sequences of travellers. To improve the understanding of the underlying mechanisms that generate such sequences, the *HypTrails* approach provides for a novel data analysis method. Based on first-order Markov chain models and Bayesian hypothesis testing, it allows for comparing a set of hypotheses, i.e., beliefs about transitions between states, with respect to their plausibility considering observed data. HypTrails has been successfully employed to study phenomena in the online and the offline world. In this talk, we want to give an introduction to HypTrails and showcase selected real-world applications on urban mobility and reading behavior on Wikipedia.

1 Introduction

Today, large collections of data are available in the form of sequences of transitions between discrete states. For example, people move between different locations in a city, users navigate between web pages on the world wide web, or users listen to sequences of songs of a music streaming platform. Analyzing such datasets can leverage the understanding of behavior in these application domains. In typical machine learning and data mining approaches, parameters of a model (e.g., Markov chains) are learned automatically in order to capture the data generation process and make predictions. However, it is then often difficult to interpret the learned parameters or to relate them to basic intuitions and existing theories about the data, specifically if many parameters are involved. In a recently introduced line of research, we therefore aim to establish an alternative approach: we develop a method that allows to capture the belief in the generation

This work summarizes a previous publication presenting the HypTrails approach [5] and three selected papers [3, 1, 2] that utilize it.

of sequential data as Bayesian priors over parameters and then compare such *hypotheses* with respect to their plausibility given observed data. In this work, we want to showcase our general approach [5], which we call *HypTrails*, and present some practical applications in various domains [3, 1, 2], i.e., sequences of visited locations derived from photos uploaded to Flickr, taxi directions in Manhattan, and navigation of readers in Wikipedia.

2 Bayesian hypotheses comparison in sequential data

For comparing hypotheses about the transition behavior in sequence data, we follow a Bayesian approach. As an underlying model, we utilize first-order Markov chain models. Such models assume a memory-less transition process between discrete states. That means that the probability of the next visited state depends only on the current one. The parameters of this model, i.e., the transition probabilities p_{ij} between the states, can be written as a single matrix.

In HypTrails, we want to compare a set of hypotheses $H_1, \ldots, H_n$ with respect to how well they can explain the generation of the observed data. Each of the hypotheses captures a belief in the transition between the states as derived from theory in the application domain, from other related datasets, or from human intuition. To specify a hypothesis, the user can express a *belief matrix*, in which a high value in a cell (i, j) reflects a belief that transitions between the states i and j are more common. With HypTrails, these belief matrices are then automatically transformed into Bayesian Dirichlet priors over the model parameters (i.e., the transition probabilities in the Markov chain). This transformation can be performed for different concentration parameters κ. A higher value of κ generates a prior that corresponds to a stronger belief in the hypothesis. For each hypothesis H_i, and each concentration parameter κ, we can then compute the marginal likelihood $P(D|H_i)$ of the data given the hypothesis. Given our model, the marginal likelihood can efficiently be computed in closed form. The higher the marginal likelihood of a hypothesis is, the more plausible it appears to be with respect to the observed data. For quantifying the support of one hypothesis over another, we utilize Bayes Factors, a Bayesian alternative to frequentists p-values, which can directly be interpreted with lookup tables [4]. For a set of hypotheses, the marginal likelihoods induce an ordering of the hypotheses with respect to their plausibility given the data. However, the plausibility of hypotheses is always only checked relatively against each other. Therefore, often a simple hypothesis is used as a baseline, e.g., the uniform hypothesis that assumes all transitions to be equally likely.

To compare hypotheses, all priors should be derived using the same belief strength κ. To make comparisons across different belief strength, HypTrails results are typically visualized as line plots, in which each line corresponds to one hypothesis. The x-axis specifies different values of the concentration parameter κ, and the y-axis describes the marginal likelihood of a hypothesis, cf. Figure 1.

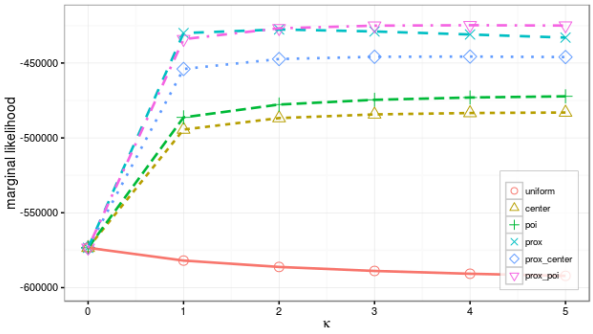

Fig. 1: **Example result of HypTrails (Flickr study, Berlin).** Each line represents one hypothesis. The x-axis defines different concentration parameters (strengths of belief), the y-axis indicates (logs of) marginal likelihoods for each hypothesis. It can be seen that the baseline "uniform" hypotheses is by far the least plausible of these hypothesis, while a mixture of proximity and center hypotheses ("prox-center") and a mixture of proximity and point-of-interest hypotheses ("prox-poi") perform best.

3 Applications

Next, we outline three real-world applications of this technique.

3.1 Urban mobility in Flickr

In a first study, we focused on geo-temporal trails derived from Flickr. In particular, we crawled all photos on Flickr with geo-spatial information (i.e., latitude and longitude) from 2010 to 2014 for four major cities (Berlin, London, Los Angeles, and New York). We used a map grid to construct a discrete state space of locations. Then, we created a sequence of locations for each user that uploaded pictures of that city based on the picture locations. On the sequences, we evaluated a variety of hypotheses such as a proximity hypothesis (next location is near the current one), a point-of-interest hypothesis (next location will be at a tourist attraction or transportation hub), a center hypothesis (next location will be close to the city center), and combinations of them. As a result, rankings are mostly consistent across cities. Combinations of proximity and point-of-interest hypotheses are overall most plausible. Figure 1 shows example results for Berlin.

3.2 Taxi usage in Manhattan

In a second study, we investigated again trails of urban mobility. In particular, we studied a dataset of taxi trails in Manhattan[5]. In this study, we used *tracts* (small administrative) units as a state space of locations. Using additional information

[5] http://www.andresmh.com/nyctaxitrips/

on these tracts extracted from census data and data from the FourSquare API, we investigated more than 60 hypotheses such as "taxis drive to tracts with similar ethnic distribution" or "taxis will drive to popular locations w.r.t. check-ins". We also performed spatio-temporal clustering of the sequence data and applied HypTrails on the individual clusters to find behavioral traits that are typical for certain times and places. For instance, we discovered a group of taxi rides to locations with a high density of party venues on weekend nights.

3.3 Link usage in Wikipedia

In another work, we studied transitions between articles in the online encyclopedia Wikipedia. In particular, we were interested in which links on a Wikipedia page get frequently used. For that purpose, we applied HypTrails on a recently published dataset of all transitions between Wikipedia pages for one month[6] using the set of all articles as state space. For constructing hypotheses, we considered hypotheses based on visual features of the links (e.g., "links in the lead paragraph get clicked more often" or "links in the main text get clicked more often"), hypothesis based on text similarity between articles, and hypotheses based on the structure of the link network of Wikipedia articles. As a result, hypotheses that assume people to prefer links at the top and left-hand side, and hypotheses that express a belief in more frequent usage of links towards the periphery of the article network are most plausible.

4 Conclusion

In this work, we gave a short introduction into the HypTrails approach that allows to compare the plausibility of hypotheses about the generation of a sequential datasets. Additionally, we described three real-world applications of this technique for studying urban mobility and reading behavior in Wikipedia.

References

1. Becker, M., Singer, P., Lemmerich, F., Hotho, A., Helic, D., Strohmaier, M.: Photowalking the city: Comparing hypotheses about urban photo trails on Flickr. In: Int. Conference on Social Informatics (SocInfo). pp. 227–244 (2015)
2. Dimitrov, D., Singer, P., Lemmerich, F., Strohmaier, M.: What makes a link successful on Wikipedia? In: Int. World Wide Web Conference. pp. 917–926 (2017)
3. Espín Noboa, L., Lemmerich, F., Singer, P., Strohmaier, M.: Discovering and characterizing mobility patterns in urban spaces: A study of Manhattan taxi data. In: Int. Workshop on Location and the Web. pp. 537–542 (2016)
4. Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association 90(430), 773–795 (1995)
5. Singer, P., Helic, D., Hotho, A., Strohmaier, M.: Hyptrails: A Bayesian approach for comparing hypotheses about human trails on the web. In: Int. World Wide Web Conference. pp. 1003–1013 (2015)

[6] https://datahub.io/dataset/wikipedia-clickstream