Modeling Dynamic Behaviors within Population

Nazgol Tavabi
USC Information Sciences Institute
nazgolta@isi.edu

Abstract

The abundance of temporal data generated by mankind in recent years gives us the opportunity to better understand human behaviors along with the similarities and differences in groups of people. Better understanding of human behaviors could be very beneficial in choosing strategies, from group-level to society-level depending on the domain. This type of data could range from physiological data collected from sensors to activity patterns in social media. Identifying frequent behavioral patterns in sensor data could give more insight into the health of a community and provoke strategies towards improving it; By analyzing patterns of behaviors in social media, platform's attributes could be adjusted to the user's needs.

This type of modeling introduces numerous challenges that varies depending on the data. The goal of my doctoral research is to introduce ways to better understand and capture human behavior by modeling individual's behaviors as time series and extracting interesting patterns within them.

Introduction

Time series data appears in many different applications from health records to industrial processes and etc, and tasks like time series classification/regression where multiple time series need to be Modeled/analyzed together remain among most important and challenging tasks (Yang and Wu 2006). Joint modeling of multiple time series is specifically beneficial when we are trying to learn attributes from a population of individuals, each person represented as a univariate/multivariate time series. This data could include anything from participant's biometric signals to their activity on social media, tone of their voice, their GPS data and etc.

Modeling multiple time series in order to understand individual's behavior introduces many challenges. For example, most models for time series analysis accept synchronous time series with fixed lengths as input. However in modeling behaviors this is rarely the case and often data consists of asynchronous data with different lengths. Another challenge in these problems is having an interpretable representation. More often than not interpretability loses in the trade-off with prediction power. Especially with the growth of deep learning methods. However with the goal of understanding patterns and behaviors in the data, interpretable representations become much more important.

The focus of my doctoral research is to model multiple time series, representing human behavior, and propose valid and interpretable representations for them and to identify and analyze dynamic behaviors within them.

Completed Research

In this section I will briefly describe two research projects I have done in analyzing dynamic behavior in the data.

Characterizing Activity on the Deep and Dark Web (Tavabi et al. 2019a) In this work we studied a large corpus of messages posted to the deep and darkweb (d2web) forums over a period of more than a year. D2web refers to limited access web sites that require registration, authentication, or more complex encryption protocols to access them. These web sites serve as hubs for a variety of illicit activities: to trade drugs, stolen user credentials, hacking tools, and to coordinate attacks and manipulation campaigns. In this work patterns and behaviors are studied at forum-level as opposed to user-level. We identify topics of discussion by running Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) on messages posted to forums. Then, use LDA's weights to represent forums as multivariate time series. Signals in the multivariate time series are different topics found by LDA and each value is calculated by averaging LDA weights of messages posted to a forum in a week. To model the evolution of topics across forums we use a non-parametric HMM (Fox et al. 2014), where each forum is modeled by a separate HMM though the states are shared among different time series. With the identified states, we can examine the dynamic patterns of discussion and cluster forums with similar patterns. Figure 1 shows the resulting dendrogram and also the sequences of learned states for each forum. Each line in the figure represents a forum, and different states are represented by different colors. Transitions between states are visible in places the colors alternate. We show that our approach surfaces hidden similarities across different forums and can help identify anomalous events in this data by taking advantage of rare states seen in the results.

Learning Behavioral Representations from Wearable Sensors (Tavabi et al. 2019b) In this work we studied biometric and movement signals collected with wearable

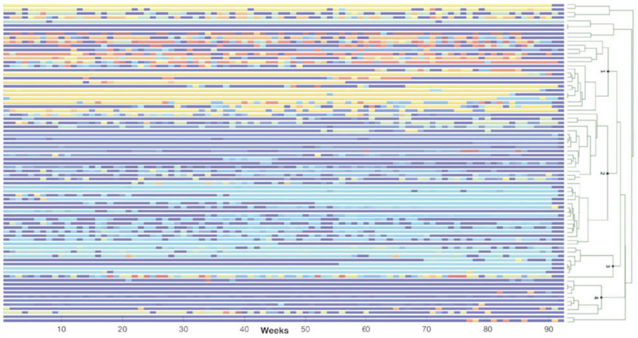

Figure 1: State sequences of forums (each line represents a forum and each color represents a state) and Dendrogram showing the similarity of forums based on their learned states.

sensors from people going on about their daily lives. We collected data from sensors worn by a group of workers in a large urban hospital and asked them to complete surveys prior to the study. Pre-study surveys measured job performance, cognitive ability, personality, affect, and health states, which serve as ground truth constructs for our study.

We used non-parametric HMM, the same model used in the previous section, to identify states, behaviors, shared among participants. We define two different distance measures between individuals based on the learned HMMs and the shares states. We also propose two representations for participants:

1. A compact representation of participants to interpret the learned latent states and gain more insight into the data.
2. A more advanced but less interpretable representation to predict individual attributes, such as personality traits, age, etc.

The first representation is the stationary distribution of each HMM which is the relative amount of time spent in each state and since states are shared among participants they can be treated as dimensions of the embedding space. In the paper we show that by using these representations as features to predict pre-study surveys, we can better understand the behavioral states. An example of it is given below:

By looking at frequency of a single state, state A, we are able to distinguish day-shift nurses with night-shift nurses. Average movement signals of state A show minimal to zero movement and horizontal alignment of the participant; we can assume that this state correspond to quick rests within shifts. State A also helps predict positive affect (POS-AF), well-being, and hindrance stress. State A has a positive coefficient in predicting POS-AF and Well-being, whereas for hindrance stress it has a negative coefficient. Hindrance stress is generally perceived as a type of stress that prevents progress toward personal accomplishments. Thus, a plausible interpretation of these results is: Quick naps or breaks during work hours could increase positive affect and well-being and decrease hindrance stress

The second representation uses one of two similarity measures proposed, to get a distance matrix of participants, computes the normalized Laplacian of the distance matrix and retrieves K largest eigen-vectors (i.e., eigen-vectors corresponding to largest eigenvalues) as representations of participants (K is a hyperparameter). This representation is not interpretable like the first one, however it outperforms the competing methods in predicting personality traits.

Two different similarity measures are proposed because they differ in their sensitivity to small changes in the data. This trade-off causes one method to perform better than the other depending on the construct we want to predict. Constructs include: positive affect, negative affect, introversion, extroversion and etc. There are 33 different target constructs.

Current Research

Currently I am working on a method for time series embedding. Recent advances of Natural Language Processing (NLP) have introduced many successful embedding techniques for text, some of these techniques could be very well adopted/adjusted for embedding time series. Moreover by identifying interpretable embeddings, we can use them to do feature selection and choose the signals relevant to the labels we want to predict, this option becomes very useful in situations where we have rich dataset with many signals such as dataset used in (Tavabi et al. 2019b) where we have heart rate, breathing rate, movement, etc; and only a subset of these signals might be relevant in predicting a construct such as stress.

Future Work

One possible direction for future work is capturing casual dependencies between time series. Within the same project mentioned in paper (Tavabi et al. 2019b) we have collected daily surveys from participants, asking their level of stress, mood, alcohol and tobacco usage, whether an atypical event has happened, etc. We are interested to identify patterns and behaviors in this data and understand the relationship among these different signals.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Fox, E. B.; Hughes, M. C.; Sudderth, E. B.; Jordan, M. I.; et al. 2014. Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *The Annals of Applied Statistics* 8(3):1281–1313.

Tavabi, N.; Bartley, N.; Abeliuk, A.; Soni, S.; Ferrara, E.; and Lerman, K. 2019a. Characterizing activity on the deep and dark web. In *Companion Proceedings of The 2019 World Wide Web Conference*, 206–213. ACM.

Tavabi, N.; Hosseinmardi, H.; Villatte, J. L.; Abeliuk, A.; Narayanan, S.; Ferrara, E.; and Lerman, K. 2019b. Learning behavioral representations from wearable sensors.

Yang, Q., and Wu, X. 2006. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making* 5(04):597–604.