Final Project Proposal: LASTS Stress-Testing Via Adversarial Datasets

Rob Schwartz, Harry Li
{res7cd,jl9dbb}@virginia.edu

March 24, 2022

1 Introduction (2')

We choose to replicate and build on an existing conference paper, Explaining Any Time Series Classifier by Riccardo Guidotti, Anna Monreale, Francesco Spinnato, Dino Pedreschi, Fosca Giannotti in CogMI 2020.

The conference paper by Guidotti et. al, poses a novel method for explaining time series models, known as LASTS (Local Agnostic Subsequence-based Time Series Explainer). This model-agnostic explanation method uses an autoencoder based system to generate representative results in the latent sample space, then trains decision trees in the latent space to provide explainable and accurate decision exemplars and rules. Given a lack of accurate counterfactual time series explanation techniques in the research literature, LASTS also improves on previous literature by offering an explanation method that (a) does not require the discretization size of the input time series as a hyperparameter, (b) provides native counterfactuals, and (c) outputs explanations in both exemplar and rule-based formats (both outputs are provided in factual and counterfactual types).

In contrast to previous SHAP-based time series explanation models—which either yield fixed-length saliency blocks or point saliency indications (similar to pixel saliency maps)—LASTS uses the shapelet as the explanatory unit. The shapelet is a more interpretable data format because it represents a maximally distinctive subsample of the time series of the class.

However, the authors of this shapelet-focused paper implicitly raise a particular criticism by electing shapelets as the LASTS unit of interest while simultaneously arguing that LASTS is model-agnostic: what happens when classification of the underlying time series dataset is poorly explained by shapelets? Will LASTS still perform well?

2 Models and datasets (1')

As LASTS is a model-agnostic method, the authors use ResNet, a CNN, and a KNN to create the black-box classification models. We aim to leave these models and their hyperparameters the same.

We also plan to test out LASTS using two of the same datasets as the authors: cylinder-ball-funnel, and epileptic seizure recognition. This result will replicate the author's analysis. Both of these datasets are well-suited to shapelet-based analysis.

Lastly, we will create two new time series datasets that exhibit poor shapelet explainability but good SHAP explainability. We will evaluate LASTS on these two datasets using the same methods as for the first two datasets. Please find more information on these new datasets below.

3 Proposed method (1')

The two new time series datasets that will be created will be named "waves" and "gaps". They are targeted to be more explainable using SHAP-based methods than LASTS.

"Waves" will be a dataset consisting of various smooth curves with a high-frequency sine wave added over portions of each sample. The ground-truth classification will be based on the percentage of the sample that contains a high-frequency portion: if above a certain threshold, then the classification will be positive: otherwise, the classification will be negative. The number of transitions from presence to absence of the high-frequency sine wave will remain indistinguishable in the positive and negative sets, so that the transitions themselves cannot be relevant to the shapelet-based explanation.

"Gaps" will be a dataset consisting of various constant measurements that change between a low value and a high value. The ground-truth classification will be based on the percentage of the sample that contains a high value: if above a certain threshold, then the classification will be positive: otherwise, the classification will be negative. The number of transitions from high to low will remain indistinguishable in the positive and negative sets, so that the transitions themselves cannot be relevant to the shapelet-based explanation.

If our new "waves" and "gaps" datasets fail to train the black-box models with good accuracy, then we can also modify the datasets to make the classes more discriminative (for example, increasing the amplitude of the high-frequency sine wave in the "waves" dataset).

4 Experiments (2')

We will train the black-box models on all four datasets, and then train LASTS and the SHAP-based explainability techniques. We aim to measure the same qualitative and quantitative metrics as measured on the original datasets when using LASTS on each new dataset; (1) visual comparison of exemplars, (2) visual comparison of shapelet rules, (3) counter-example usefulness measurement (K-NN testing), (4) surrogate faithfulness, (5) explanation stability.

We hypothesize that the LASTS explanation method will do relatively poorly on each metric under the "waves" dataset, due to (1) the positioning of the shapelet explanation, which locates the explanation at the point of best fit on the exemplar time series (although facing what is a repeating pattern), and (2) the underlying issue that the presence or absence of the pattern alone does not justify the classification. We expect the SHAP-based explanation methods to perform better, due to their focus on portions, not shapes, of the time series that are relevant to the explanation (although they may also fail).

We hypothesize that the LASTS method will do very poorly on each metric under the "gaps" dataset, due to the complete irrelevancy of the local shape of the time series to the classification. We expect the SHAP-based explanations to perform better, due to the same reasons as stated above.

We are very excited to complete this experiment, and we truly believe that it would be a good fit for the Workshop on Insights from Negative Results in NLP if this workshop were focused on time series!