ES-Mask: Evolutionary Strip Mask for Explaining Time Series Prediction (Student Abstract)

Yifei Sun[1], Cheng Song[1], Feng Lu[1], Wei Li[2], Hai Jin[1], Albert Y. Zomaya[2]

[1] National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, China

[2] Centre for Distributed and High Performance Computing, School of Computer Science, The University of Sydney, Australia

{sunyifei,songch,lufeng,hjin}@hust.edu.cn,{weiwilson.li,albert.zomaya}@sydney.edu.au

Abstract

Machine learning models are increasingly used in time series prediction with promising results. The model explanation of time series prediction falls behind the model development and makes less sense to users in understanding model decisions. This paper proposes ES-Mask, a post-hoc and model-agnostic evolutionary strip mask-based saliency approach for time series applications. ES-Mask designs the mask consisting of strips with the same salient value in consecutive time steps to produce binary and sustained feature importance scores over time for easy understanding and interpretation of time series. ES-Mask uses an evolutionary algorithm to search for the optimal mask by manipulating strips in rounds, thus is agnostic to models by involving no internal model states in the search. The initial experiments on MIMIC-III data set show that ES-Mask outperforms state-of-the-art methods.

Introduction

The transparency of how machine learning models make decisions is vital in fostering trustworthy AI systems. Saliency methods were proposed to elucidate the bonding of the features in inputs and prediction making to help build trust among stakeholders. These methods have been implemented as explainable AI tools and worked well on images, text, and tabular data. Those tools are good at predicting the important score of a feature at a single time point (as shown in Figure 1(a)−(d)) but often ignore the time ordering of inputs and the time-sensitive nature of time series applications. Dynamask (Crabbé and Van Der Schaar 2021) was the first perturbation-based saliency method leveraging the idea of the mask to generate post-hoc explanations for any machine learning models on time series applications. The perturbation-based method uses input to produce a data variation to examine feature importance rather than variating features on the same input. The mask perturbs the time series input to measure the importance of the features to the prediction. For simplicity, the mask values are normalized between 0 and 1, where close to 1 indicates the feature is salient, 0 otherwise. This simplicity derives from the fact that explaining long time series with the results like Figure 1(a)−(d) is unlikely to be effective in practice. However, as shown in Figure 1(e), the explanation produced by Dynamask is time series friendly as it can deliver near binary mask values for cherry-picking salient features over time. It is also sobering to note that the feature saliency over time is still in a somewhat fragmented appearance. We argue that the explanation is more user-friendly if the saliency could be as consistent as possible over time. For example, the explanation as Figure 1(f) allows users to quickly capture the salient features over time by color and shape.

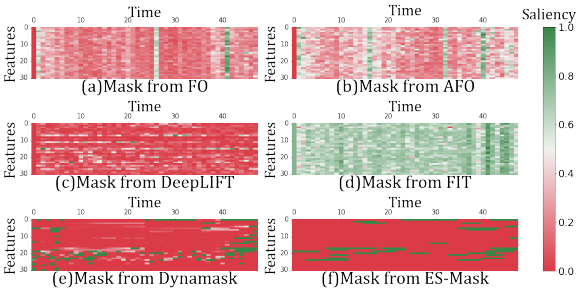

Figure 1: Explanation masks for a patient in MIMIC-III with six different methods.

In this work, we go one step further to develop an evolutionary strip mask-based saliency method (ES-Mask) to enable easy and intuitive explanations to users. ES-Mask is a post-hoc and model-agnostic approach for time series applications. The novelty of ES-Mask lies in the masks being initialized as strips than a grid of discrete points to perturb the inputs. The masks can be iteratively refined by the translation, mutation, and crossover operators until producing the best possible one for greater transparency and trust building.

Methodology

Mask and Strips

ES-Mask randomly initializes a mask set $\{\mathbf{M_i} = (m^i_{t,r}) \in \{0,1\}^{T\times d_X}\}$ on the input data $\mathbf{X} = (x_{t,r})_{(t,r)\in[1:T]\times[1:d_X]}$. Here, $i \in [1, I]$, I is the number of initialized masks, T is the duration of $\mathbf{X}$, and d_X is the number of features in $\mathbf{X}$. When $m^i_{t,r} = 1$, it means that $x_{t,r}$ is salient for the prediction, or false otherwise. In ES-Mask, mask $\mathbf{M_i}$ has a strip set $\mathbf{S_i} = \{S^i_n \mid n \in [1, N]\}$, where $N \in \mathbb{N}$ is the number of strips in a mask. For generating a strip $S^i_n \in \mathbf{S_i}$, we randomly initialize its start $b \in [1, T]$, the feature index $d \in [1, d_X]$,

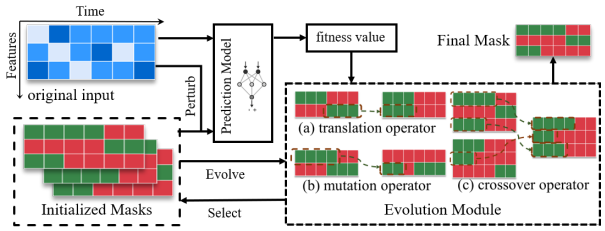

Figure 2: Overview of ES-Mask.

and the strip size l. For each point $m^i_{t,r}$ in the strip, $m^i_{t,r} = 1\ for\ all\ t \in [b, b+l]$. To evaluate the effect of the mask, we perturb the raw input with the mask. Formally, the result of perturbed input is $\Gamma_{\mathbf{M}}(\mathbf{X})$, where $\Gamma_{\mathbf{M}} : \mathbb{R}^{T\times d_X} \rightarrow \mathbb{R}^{T\times d_X}$. Assuming the prediction model is $\mathbf{Y} = f(\mathbf{X})$, the prediction of perturbed input will be $f(\Gamma_{\mathbf{M}}(\mathbf{X}))$.

We aim to maximize the squared error between the unperturbed and the perturbed predictions by ES-Mask to reveal salient features strongly correlated to the prediction. The objective function is defined as:

$$\mathbf{M}^* = \underset{\mathbf{M}\in\{0,1\}^{T\times d_X}}{\arg\max} \left(f(\mathbf{X}) - f(\Gamma_{\mathbf{M}}(\mathbf{X}))\right)^2 \quad (1)$$

Mask Evolution

Considering the ground truth explanation is not known beforehand, ES-Mask adopts the evolutionary algorithm to optimize strip set $\mathbf{S}$ in rounds to let mask $\mathbf{M}$ be the best possible answer. The overview of ES-Mask is given in Figure 2. In ES-Mask, we treat a mask as an individual and each strip as the atomic unit of optimization. We initialize a group of strips in a mask to perturb the inputs. The perturbed input and the original input will use the prediction model to evaluate the effect of the perturbation by the fitness value measured by $\left(f(\mathbf{X}) - f(\Gamma_{\mathbf{M}}(\mathbf{X}))\right)^2$ to select the masks of next generation. Besides the fitness value, we also use three operators to let the next generation mask be a better fit. The translation operator is employed to adjust the position offset of the strips on the timeline like Figure 2(a). The mutation operator (Figure 2(b)) encourages genetic diversity to help evolution and prevents ES-Mask from converging to a local optimal by emulating mutation in nature, replacing an old strip with a new strip in a generation. We also adopt the crossover operator to allow the next generation mask to inherit any strips from its parents like Figure 2(c). The optimal mask will be found after certain iterations.

Experimental Evaluation

We developed an RNN model to predict the mortality of the patients who stayed in critical care units recorded in the MIMIC-III data set. The data selection, preprocessing, and model training were the same as Dynamask. We also employed FIT (Tonekaboni et al. 2020), DeepLIFT (Shrikumar, Greenside, and Kundaje 2017), FO, and AFO as performance benchmarks in our experiment. We used mask entropy (Crabbé and Van Der Schaar 2021) (lower is better), continuity (Mercier et al. 2022) (lower is better), and cross entropy (CE, higher is better) to evaluate the performance of all approaches. Table 1 shows the experimental results. It is easy to see that ES-Mask outperformed all benchmarks in all aspects. The mask entropy of ES-Mask achieved 0 means its polarization reaches the theoretical minimum, demonstrating that the explanation of ES-Mask is binary only while the others are multiple values based and exhibit uncertainty on feature saliency. The lowest continuity indicates that ES-Mask has minimal change and the saliency is more consistent over time. This result reflects the main design difference as ES-mask initializes the mask by strips rather than a grid of points in other approaches, thus making the result closer to the ground truth. The highest CE value demonstrates that ES-Mask is most sensitive to the model prediction reacting to the perturbed input as the design of operators incorporates the time ordering and the time-sensitive nature of the input.

Methods	Mask Entropy↓	Continuity↓	CE↑
ES-Mask	**0.00**	**45.74**	**0.088**
Dynamask	0.77	98.11	0.072
FIT	26.33	144.34	0.034
DeepLIFT	128.65	113.71	0.033
FO	191.90	124.73	0.035
AFO	189.18	137.54	0.034

Table 1: Results of various methods to explain RNN used to predict the mortality rate of patients in MIMIC-III data set.

Conclusion

We proposed ES-Mask to provide a model agnostic and easy-to-understand explanation for time series applications toward improving trust. ES-Mask works well on clinical data, and we plan to demonstrate its applicability on more time series applications.

Acknowledgments

This work is supported by the Key Project of the National Natural Science Foundation of China (62232012) and the Hubei Big Data Analysis Platform and Intelligent Service Project for Medical and Health.

References

Crabbé, J.; and Van Der Schaar, M. 2021. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning*, 2166–2177. PMLR.

Mercier, D.; Bhatt, J.; Dengel, A.; and Ahmed, S. 2022. Time to Focus: A Comprehensive Benchmark Using Time Series Attribution Methods. *arXiv preprint arXiv:2202.03759*.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, 3145–3153. PMLR.

Tonekaboni, S.; Joshi, S.; Campbell, K.; Duvenaud, D. K.; and Goldenberg, A. 2020. What went wrong and when? Instance-wise feature importance for time-series black-box models. *Advances in Neural Information Processing Systems*, 33: 799–809.