Attention-Based Counterfactual Explanation for Multivariate Time Series

Chapter · August 2023

DOI: 10.1007/978-3-031-39831-5_26

4 authors:

Peiyu Li
Utah State University
12 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Omar Bahri
Utah State University
13 PUBLICATIONS **33** CITATIONS
SEE PROFILE

Soukaina Filali Boubrahimi
Utah State University
60 PUBLICATIONS **381** CITATIONS
SEE PROFILE

Shah Muhammad Hamdi
Utah State University
43 PUBLICATIONS **241** CITATIONS
SEE PROFILE

Attention-based Counterfactual Explanation for Multivariate Time Series

Peiyu Li, Omar Bahri, Soukaïna Filali Boubrahimi, and Shah Muhammad Hamdi

Department of Computer Science, Utah State University, Logan, UT 84322, USA
{peiyu.li, omar.bahri, soukaina.boubrahimi, s.hamdi}@usu.edu

Abstract. In this paper, we propose Attention-based Counterfactual Explanation (AB-CF), a novel model that generates post-hoc counterfactual explanations for multivariate time series classification that narrow the attention to a few important segments. We validated our model using seven real-world time-series datasets from the UEA repository. Our experimental results show the superiority of AB-CF in terms of validity, proximity, sparsity, contiguity, and efficiency compared with other competing state-of-the-art baselines.

Keywords: EXplainable Artificial Intelligence (XAI) · Counterfactual Explanation · Multivariate Time Series · Attention based.

1 Introduction

Over the past decade, artificial intelligence (AI) and machine learning (ML) systems have achieved impressive success in a wide range of applications. The challenge of many state-of-art ML models is a lack of transparency and interoperability. To deal with this challenge, the EXplainable Artificial Intelligence (XAI) field has emerged. A lot of efforts have been made to provide post-hoc XAI for image, vector-represented data, and univariate time series data while significantly less attention has been paid to multivariate time series data [2]. The high dimensional nature of multivariate time series makes the explanation models one of the most challenging tasks [4]. In this work, we propose a model-agnostic counterfactual explanation method for multivariate time series data. According to the recent literature on counterfactual explanations for various data modalities [4], an ideal counterfactual explanation should satisfy the following properties: validity, proximity, sparsity, contiguity, and efficiency. Our method is designed to balance these five optimal properties.

2 Related work

In the post-hoc interpretability paradigm, the counterfactual explanation method proposed by Wachter et al. [10] and the native guide counterfactual (NG-CF) XAI method proposed by Delaney et al. [4] are the two most popular methods. The first one aims at minimizing a loss function to encourage the counterfactual to change the decision outcome and keep the minimum Manhattan distance from the original input instance.

Based on this method, several optimization-based algorithms that add new terms to the loss function to improve the quality of the counterfactuals were proposed as an extension [5], [6]. NG-CF uses Dynamic Barycenter (DBA) averaging of the query time series and the nearest unlike neighbor from another class to generate the counterfactual instance. Recently, several shapelet-based and temporal rule-based counterfactual explanation methods have been proposed to provide interpretability with the guide of mined shapelets or temporal rules [7], [2], [3]. However, these aforementioned counterfactual explanation methods suffer from generating counterfactuals that are low sparsity, low validity, or high cost of processing time to mine the shapelets or temporal rules. To deal with these challenges, in our work, we propose to focus on minimum discriminative contiguous segment replacement to generate more sparse and higher-validity counterfactuals efficiently.

3 Methodology

3.1 Notation

We define a multivariate time series dataset $\boldsymbol{D} = \{\boldsymbol{X}_0, \boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ *as a collection of n multivariate time series where each multivariate time series has mapped to a mutually exclusive set of classes* $\boldsymbol{C} = \{c_1, c_2, ..., c_n\}$*. A multivariate time series classification model* f *is pretrained using the dataset* $\boldsymbol{D}$*. Dataset* $\boldsymbol{D}$ *comes with a pre-defined split ratio. We train each model only on the training data, then explain their predictions for all test instances. We define the instances from the test set as query instances. For each query instance that is associated with a class* $f(\boldsymbol{X}_i) = c_i$*, a counterfactual explanation model* $\mathcal{M}$ *generates a perturbed sample with the minimal perturbation that leads to* $f(\boldsymbol{X}'_i) = c'_i$ *such that* $c_i \neq c'_i$*. We define the perturbed sample* $\boldsymbol{X}'_i$ *as the counterfactual explanation instance, and* c'_i *as the target class we want to achieve for* $\boldsymbol{X}'_i$*.*

3.2 Proposed method

In this section, we introduce our proposed method in detail. Fig. 1 shows the process of the proposed method.

For each query instance, to generate its counterfactual explanation, we try to discover the most important top k segments at the very first step. Specifically, we use a sliding window to obtain the subsequences of the query instance, each subsequence will be considered a candidate segment. For our case of multivariate time series data, after applying a sliding window to an input multivariate time series with d dimensions $\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^d]$, we can obtain a set of all candidates segments as follows:

$$subsequences = \{\{x_i^1, x_{i+1}^1, ..., x_{i+L-1}^1\}, ..., \{x_i^d, x_{i+1}^d, ..., x_{i+L-1}^d\}\}, \quad (1)$$

where $1 \leq i \leq m - L + 1$ is the starting position of the sliding window in the time series and $L = 0.1 \times m$ is the sliding window width, stride = L.

After we obtain the set of candidate segments, we fit the $subsequences$ to a pretrained model f. It is noteworthy that f is a pre-trained model on the training set data of

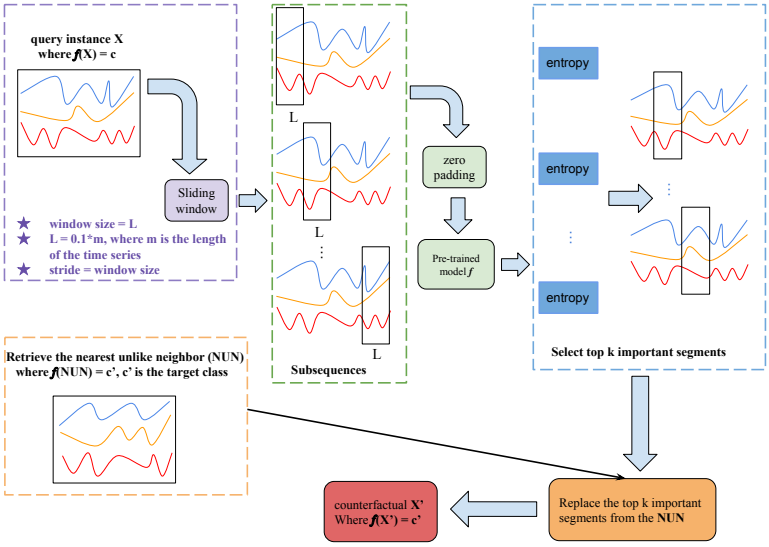


Fig. 1: Proposed Counterfactual Explanation Method

full length, therefore we used the padding technique to concatenate the segments with zeroes until the length is equal to the full length of the original time series data.

Next, each candidate is predicted by the pre-trained model, and the outcome is a probability vector $[p_1, ..., p_n]$ for the classification of n classes. We use Shannon entropy, introduced by Shannon [9], as listed in Equation 2 to measure the information embedded in each segment given the probability distribution.

$$E = -\sum_{c=1}^{|n|} p_c log_2(p_c) \geq 0 \tag{2}$$

where p_c is the class probability of class c the pre-trained model f predicted and $|n|$ is the total number of classes.

The information entropy E determines whether the candidate is discriminative. According to the information theory [9], if the information entropy is maximized, indicates that the prediction probabilities tend to follow a uniform distribution, which further indicates that this candidate fails to provide discriminative information for the model to make a prediction. On the opposite, the closer E to 0, the more important information that the candidate subsequence carries for the model to make a prediction. Consequently, we selected the top k distinguishable segment candidates with the lowest values of information entropy. k is a parameter representing the number of focused parts to be considered in the proposed method. k is initialized to 1 and increased during the counterfactual explanation generation until the generated counterfactual explanation satisfies the validity property. In the end, k will be determined as the minimum value that can make sure the counterfactual instance is classified to the target class. The final step is

Table 1: UEA datasets metadata

ID	Dataset Name	TS length	DS train size	DS test size	Dimensions	classes
0	ArticularyWordRecognition	144	275	300	9	25
1	BasicMotions	100	40	40	6	4
2	Cricket	1197	108	72	6	12
3	Epilepsy	206	137	138	3	4
4	ERing	65	30	270	4	6
5	NATOPS	51	180	180	24	6
6	RacketSports	30	151	152	6	4

to substitute the top k segments from the nearest unlike neighbor. This nearest unlike neighbor is the 1-nearest neighbor from the target class in the training dataset.

4 Experiments

4.1 Datasets

We evaluated our proposed method on the publicly-available multivariate time series data sets from the University of East Anglia (UEA) MTS archive [1]. In particular, we selected seven datasets from the UEA archive that demonstrate good classification accuracy on state-of-the-art classifiers as reported in [8]. This ensures the quality of our generated counterfactual instances. Table 1 shows the metadata of the seven datasets.

4.2 Baseline Methods

We evaluated our proposed method with the following two baselines:

- **Native guide counterfactual (NG-CF):** NG-CF uses Dynamic Barycenter (DBA) averaging of the query time series **X** and the nearest unlike neighbor from another class to generate the counterfactual example [4].
- **Alibi Counterfactual (Alibi):** The Alibi follows the work of Wachter et al. [10], which constructs counterfactual explanations by optimizing an objective function,

$$L = L_{pred} + \lambda L_{dist}, \tag{3}$$

4.3 Experimental result

In this section, we utilize the following four evaluation metrics to compare our proposed method with the other two baselines concerning the desirable properties of counterfactual instances according to the literature review.

The first one is *target probability*, which is used to evaluate the validity property. We define the **validity** metric by comparing the target class probability for the prediction of the counterfactual explanation result. The closer the target class probability is to 1, the better. The second one is the *L1 distance*, which is used to evaluate the **proximity** property. We measure the L1 distance between the counterfactual instance and the

query instance, a smaller L1 distance is desired. Then we calculate the percent of data points that keep unchanged after perturbation to show the **sparsity** level. A high sparsity level that is approaching 100% is desirable, which means the time series perturbations made in $\mathbf{X}$ to achieve $\mathbf{X}'$ is minimal. Finally, we compare the *running time* of the counterfactual instances generation to verify the **efficiency**, the faster a valid counterfactual instance can be generated the better.

Fig 2 shows our experimental results, the error bar shows the mean value and the standard deviation over each dataset set. From Fig 2b, we note that our proposed method achieves a competitive L1 distance compared with NG-CF. For the ALIBI method, even though it achieves the lowest L1 distance, the counterfactuals' target probability achieved by ALIBI is lower than 50% (see Fig 2a), which means the counterfactuals generated by ALIBI are not even valid. In addition, from Fig 2d, we can notice that the running time of using ALIBI is much higher than our proposed method and NG-CF. In contrast, our proposed method can achieve the highest target probability within one second. Our approach also demonstrates clear advantages in terms of sparsity compared to the two baselines (see Fig 2c). This is because we only replace a few segments during perturbation, while almost half of the data points remain unchanged. In summary, our proposed method generates counterfactual explanations that balance the five desirable properties while NG-CF and ALIBI tend to maximize one property with the cost of compromising the others. The source code of our model and more visualization results figures are available on our AB-CF project website [1].

5 Conclusion

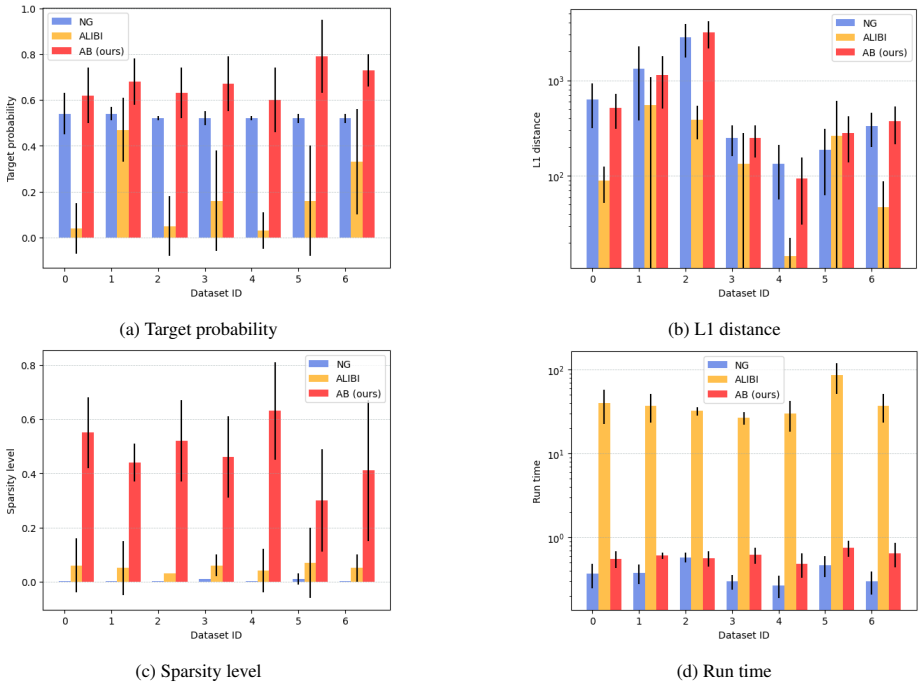
In this paper, we propose to extract the most distinguishable segments from high-dimensional data and only focus on those distinguishable segments during perturbation to avoid changing the whole time series data to obtain valid counterfactuals. Our experiments demonstrate that our proposed method stands out in generating counterfactuals that balance the validity, sparsity, proximity, and efficiency compared with the other baselines. Our method also shows a high level of contiguity since only several single contiguous segments need to be replaced.

Acknowledgments This project has been supported in part by funding from GEO Directorate under NSF awards #2204363, #2240022, and #2301397 and the CISE Directorate under NSF award #2305781.

References

1. Bagnall, A., Dau, H.A., Lines, J., Flynn, M., Large, J., Bostrom, A., Southam, P., Keogh, E.: The uea multivariate time series classification archive, 2018. arXiv preprint arXiv:1811.00075 (2018)
2. Bahri, O., Boubrahimi, S.F., Hamdi, S.M.: Shapelet-based counterfactual explanations for multivariate time series. arXiv preprint arXiv:2208.10462 (2022)

[1] https://sites.google.com/view/attention-based-cf



(a) Target probability

(b) L1 distance

(c) Sparsity level

(d) Run time

Fig. 2: Comparing the performances of NG, ALIBI, and AB models in terms of L1 distance, sparsity, target probability, and run time

3. Bahri, O., Li, P., Boubrahimi, S.F., Hamdi, S.M.: Temporal rule-based counterfactual explanations for multivariate time series. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 1244–1249. IEEE (2022)
4. Delaney, E., Greene, D., Keane, M.T.: Instance-based counterfactual explanations for time series classification. In: International Conference on Case-Based Reasoning. pp. 32–47. Springer (2021)
5. Filali Boubrahimi, S., Hamdi, S.M.: On the mining of time series data counterfactual explanations using barycenters. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 3943–3947 (2022)
6. Li, P., Bahri, O., Boubrahimi, S.F., Hamdi, S.M.: Sg-cf: Shapelet-guided counterfactual explanation for time series classification. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 1564–1569. IEEE (2022)
7. Li, P., Boubrahimi, S.F., Hamd, S.M.: Motif-guided time series counterfactual explanations. arXiv preprint arXiv:2211.04411 (2022)
8. Ruiz, A.P., Flynn, M., Large, J., Middlehurst, M., Bagnall, A.: The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery **35**(2), 401–449 (2021)
9. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE mobile computing and communications review **5**(1), 3–55 (2001)
10. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)