Counterfactual Explanations with Domain Knowledge in Multivariate Time Series

Emmanuel C. Chukwu[1], Rianne M. Schouten[1], Monique Tabak[2], and Mykola Pechenizkiy[1]

[1] Eindhoven University of Technology, Netherlands
[2] University of Twente, Netherlands
{e.c.chukwu}@tue.nl

Introduction: Counterfactual explanations (CFEs) offer an intuitive way to understand complex machine learning models by generating hypothetical scenarios that show how changes in inputs affect outcomes [6,9,10]. However, their application to multivariate time series (MTS) data remains limited [7]. Current methods often produce unrealistic or impractical counterfactuals by failing to take into account domain-specific constraints and dependencies [2, 5, 11]. MTS data are inherently complex due to interdependent variables that change over time, making theoretically valid counterfactuals difficult to implement. For example, a CFE may suggest lowering a patient's blood pressure to improve prognosis, but this may not be feasible if the patient's condition or medications prevent it. These methods use proxies for plausibility to avoid infeasible recommendations [1], many overlook the need for actionability and alignment with domain constraints [2, 8].

This paper enhances CFEs by incorporating domain knowledge. We apply an attention-based algorithm to diabetes patient data, generating counterfactuals related to physical activity to improve glucose control. Preliminary results show that our method produces actionable and feasible counterfactuals.

Background: The Attention Based Counterfactual (AB-CF) method [8] generates CFEs for MTS by identifying key data segments and replacing them with the most similar, but contrasting, segments from the training set. However, there are several downsides to this approach. First, it generates only a single CFE per instance, leaving users without multiple options to choose from. Second, in practice, the method may need to be tailored to specific data types and existing knowledge, which limits its immediate applicability in such situations. Additionally, the method relies on the "nearest unlike neighbor" mechanism, substituting parts of the instance being explained, which could further restrict its real-world use when training data are inaccessible due to privacy or ethical concerns.

Proposed Solution: We aim to incorporate domain knowledge into the generation of CFEs to create actionable and feasible counterfactuals. Domain knowledge can be formalized as rules, constraints, and expertise specific to a particular field. In this paper, we incorporate domain knowledge into the CFE generation process by identifying which univariate time series in a multivariate set are actionable, focusing on a real-world use case: diabetes management using data from the Dialect study. This observational study involves patients with type 2 diabetes treated at Ziekenhuis Groep Twente hospitals in Almelo and Hengelo [3]. We collect data on patient step count, heart rate, and blood glucose levels.

Analysis of the dataset revealed that the step count is the only modifiable variable among the three time series variables, and it follows a half-normal distribution [4]. Our proposed solution involves three key steps: 1. we select the modifiable time series. In the Dialect study, these are the step count data. 2. We incorporate knowledge of the half-normal distribution into the algorithm to model the distribution of step count data. The algorithm selects 5-length segments from the step count data and perturbs them with random values drawn from the half-normal distribution to generate counterfactuals, ensuring realistic changes. 3. Instead of creating a single CFE per instance, we generate a diverse set of CFEs using the dynamic time warping (DTW) distance. DTW measures the similarity between two TS data through an optimal alignment between them.

Results: The dataset consisted of 80 patients, with time series data recorded over two weeks. After preprocessing, the data was split into 150-minute-long subsequences, resulting in more than 80,000 instances for all patients. The dataset was then split at the patient level into training and testing sets in an 80:20 ratio. An LSTM model, consisting of two LSTM layers with 50 units each, was trained over 50 epochs with a batch size of 64 to classify glucose levels into four categories: hypoglycemia, normoglycemia, prediabetes, and hyperglycemia. A patient predicted to be in the range of hyperglycemic, hypoglycemia, or prediabetes was selected to generate CFE aimed at normoglycemia. For example, Table 1 shows an example of the CFEs generated for one instance. In this result, five segments of the data instance were perturbed, as indicated by the numbers in blue in the table. The results suggest minimal changes for the patient; however, the CFE recommends a variety of stepping activities for the subject.

Conclusion: In conclusion, the lack of actionability and feasibility in current methods for generating CFEs in MTS can be addressed by incorporating domain knowledge. This approach ensures that counterfactuals respect reality constraints, making them more practical and useful for decision-making. This work lacks a vital factor in glucose control, namely food intake, and a rigorous evaluation of the approach. Future work will adapt this approach to other CFE methods and evaluate their performance using qualitative and quantitative metrics, to provide actionable insights for time-series-based decision-making applications.

Table 1: Step count CFEs: changing patient's glucose level from prediabetes to normal glucose range

original	41	7	164	52	151	19	27	146	251	0
CFE_1	41	7	164	52	102	161	13	149	54	0
CFE_2	41	9	29	78	4	94	27	146	251	0
CFE_3	41	16	270	475	187	228	27	146	251	0
CFE_4	353	152	388	148	44	19	27	146	251	0

References

1. Asemota, A., Hooker, G.: Longitudinal counterfactuals: Constraints and opportunities (2024), https://arxiv.org/abs/2403.00105
2. Ates, E., Aksar, B., Leung, V.J., Coskun, A.K.: Counterfactual Explanations for Multivariate Time Series. In: 2021 International Conference on Applied Artificial Intelligence (ICAPAI). pp. 1–8 (May 2021). https://doi.org/10.1109/ICAPAI49758.2021.9462056, https://ieeexplore.ieee.org/document/9462056
3. den Braber, N., Vollenbroek-Hutten, M.M.R., Oosterwijk, M.M., Gant, C.M., Hagedoorn, I.J.M., van Beijnum, B.J.F., Hermens, H.J., Laverman, G.D.: Requirements of an Application to Monitor Diet, Physical Activity and Glucose Values in Patients with Type 2 Diabetes: The Diameter. Nutrients **11**(2), 409 (Feb 2019). https://doi.org/10.3390/nu11020409, https://www.mdpi.com/2072-6643/11/2/409, number: 2 Publisher: Multidisciplinary Digital Publishing Institute
4. Cooray, K., Ananda, M.: A generalization of the half-normal distribution with applications to lifetime data. Communications in Statistics-theory and Methods - COMMUN STATIST-THEOR METHOD **37**, 1323–1337 (03 2008). https://doi.org/10.1080/03610920701826088
5. Delaney, E., Greene, D., Keane, M.T.: Instance-Based Counterfactual Explanations for Time Series Classification. In: Sánchez-Ruiz, A.A., Floyd, M.W. (eds.) Case-Based Reasoning Research and Development. pp. 32–47. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-86957-1_3
6. Guidotti, R.: Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery (Apr 2022). https://doi.org/10.1007/s10618-022-00831-6, https://doi.org/10.1007/s10618-022-00831-6
7. Höllig, J., Kulbach, C., Thoma, S.: TSInterpret: A unified framework for time series interpretability (Aug 2022), http://arxiv.org/abs/2208.05280, arXiv:2208.05280 [cs]
8. Li, P., Bahri, O., Boubrahimi, S.F., Hamdi, S.M.: Attention-based counterfactual explanation for multivariate time series. In: Wrembel, R., Gamper, J., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) Big Data Analytics and Knowledge Discovery. pp. 287–293. Springer Nature Switzerland, Cham (2023)
9. Theissler, A., Spinnato, F., Schlegel, U., Guidotti, R.: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions. IEEE Access **10**, 100700–100724 (2022). https://doi.org/10.1109/ACCESS.2022.3207765, https://ieeexplore.ieee.org/document/9895252, conference Name: IEEE Access
10. Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., Shah, C.: Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. ACM Computing Surveys p. 3677119 (Jul 2024). https://doi.org/10.1145/3677119, https://dl.acm.org/doi/10.1145/3677119
11. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR (Mar 2018). https://doi.org/10.48550/arXiv.1711.00399, http://arxiv.org/abs/1711.00399, arXiv:1711.00399 [cs]