An Unsupervised Framework for Semantics Driven Causal Explanations for Anomalies[*]

Bhanukiran Vinzamuri, Elham Khabiri, and Anuradha Bhamidipaty

IBM T.J. Watson Research Center, Yorktown Heights, USA
Bhanu.Vinzamuri@ibm.com
{khabiri,anubham}@us.ibm.com

Abstract. Explainability for anomaly detection from multivariate time series sensor data procured from large assets in Industry 4.0 is a challenging and relatively unexplored problem. Apart from the temporal nature of time series data itself, another challenging aspect of this problem is the necessity of making the explainer context aware by infusing semantics which may originate from a different data modality. To address this problem, we present a workflow for the first-of-its-kind semantics-driven causal explainer for time series which uses Bayesian Network Structure learning techniques and tailors them for the anomaly explanation problem by simultaneously leveraging the knowledge of asset semantics from a graph model. We also present explanatory insights obtained from investigating a mechanical vibration anomaly for a steam turbine which was validated in our engagement with a large European energy company.

1 Introduction

The need for explainable AI has become more evident with the rise of deep learning techniques which are often accurate in practice, but are very opaque black boxes and do not offer any interpretable insights. Site Reliability Engineer (SRE) monitoring large assets are often interested in obtaining *causal* explanations which can identify cause effect relationships and are also known to be robust against identifying spurious correlations. Causal explanations can be made more meaningful in practice if they are obtained over semantically relevant features which can be validated by the SRE more effectively. In this paper, we present an approach to obtain causal local explanations for anomalies observed in large power plant assets (such as steam turbines) using semantics in conjunction with Bayesian Network Structure learning to capture causation rather than correlation. One of the key novelties of our approach is that it can obtain local explanations by learning from multivariate time series sensor data in an unsupervised fashion. This approach does not apriori need a black box output to be explained making it different from traditional black box explainability. However, the proposed method can be modified to explain black box outputs also if needed. Semantic graph models play a vital role in our framework which encode the structural model of the asset describing the relationships among different components. Each physical component maps to a set of sensors, work orders and notifications. The semantically relevant sensors (a pool of candidate global explanations) for a potential anomaly is leveraged by our causal explainer to identify the relevant sensors to explain an anomaly.

2 Background

We now provide some background to readers on how to estimate and compare causal graphs from noisy non-linear time series data. Most Bayesian Network-based graph learning methods suffer from computational complexity issues as the number of nodes increases which makes it more important to use scalable heuristic-based techniques. So, we develop a causal Bayesian Network technique which uses scalable greedy hill-climbing and mutual information based non-linear directed information testers to estimate directionality (cause-effect). Subsequently, nodes in these graphs can be compared using metrics such as Hamming distance which compares the adjacency matrices.

3 Proposed Method

We now describe the steps involved in our approach.

- SRE identifies a time window of interest based on visual inspection of multivariate sensor data to investigate an anomaly looking for an explanation.
- The semantic graph model is queried to identify a global explanation consisting of candidate sensors which correspond to the component failure.
- Multiple causal Bayesian network graphs are inferred successively using method described above over uniform periodically sampled time intervals across the entire window of interest.
- Graphs which encode the causal mechanism over each window are compared successively as explained above to identify the time of onset of the failure corresponding to time point causing maximum graph structural change. The graphs preceding and succeeding this point of onset explain the anomaly.

4 Results

We applied our approach outlined above for explaining a critical mechanical vibration anomaly in a steam turbine. Mechanical vibration node from the semantic graph model provided us with the candidate list of 40 sensors for the causal explainer to identify local explanations. SRE investigated a 60 day window preceding the date of critical failure. Our approach was able to identify, a) the point of onset of the failure (approximately 40 days preceding the critical failure), and b) journal bearing 4 was identified as the sensor with highest Hamming distance before and after onset which was validated by the SRE after inspecting the post failure repair logs. Our approach is agnostic to the kind of anomaly being explained and can be used for explaining other kinds of critical failures also. Scalability of our approach can be further improved by pre-computing the causal graphs in the backend over the entire time horizon allowing the SRE to investigate multiple windows efficiently.