CAREER: Explainable Anomaly Detection: from Association to Causation

Overview

Anomaly detection, which aims to find samples that deviate from normal ones, has attracted much attention because of a wide spectrum of applications, such as transaction fraud detection, system malfunction detection, and intrusion detection. As many deep anomaly detection models are deployed for high-stakes tasks, it is hard to trust the automatic decision by only providing a simple alarm without any evidence. Therefore, achieving explainable anomaly detection is fundamental to building a trustworthy anomaly detection system in real-world. This project aims to develop explainable anomaly detection models. We first categorize the need for explainability in security applications in three tiers: 1) Explainable -- why a sequence is labeled as abnormal; 2) Diagnostic -- what is the root cause of an anomaly 3) Actionable -- how to mitigate the abnormal status; 3) Trustworthy -- whether we can trust the explanations. Accordingly, we outline four research components: 1) developing post hoc explanation approaches that can highlight abnormal features in an anomaly; 2) developing the root cause identification approaches to locate the root cause features with the consideration of underlying causal relationships between features; 3) developing actionable explanation approaches that can recommend actions to fix the abnormal status; 4) developing robust and fair explanations to ensure the trustworthiness of explanations. The project will also develop an open-source package to benefit both the research community and the industry.

Intellectual Merit

The intellectual focus of this project is centering around the questions the domain users could ask in anomaly detection scenarios: “why is an anomaly; what causes it; how to fix it; and can we trust the explanations?” Along this line, this project explores the design of explainable models that can improve the trustworthiness of anomaly detection tasks and has the following intellectual contributions: 1) explainable anomaly detection models with counterfactual explanations; 2) diagnostic explanations models that can localize the root cause of the abnormal status by considering causal relationships between features; 3) actionable explanation models that can further recommend actions based on the root cause features to fix the abnormal status; 4) trustworthy explanations ensuring the robustness and fairness of explanations. The proposed framework meets the urgent need for explanations of highly sensitive anomaly detection tasks.

Broader Impacts

As anomaly detection tasks are commonly used in various high-stakes domains, such as cybersecurity, health, and finance, this project has the potential to lay a foundation for industries and governments to build trustworthy anomaly detection models by not just detecting abnormal behaviors but also providing explanations. The proposed research will significantly advance the explainability techniques for anomaly detection. The findings, tools, software code, and documents will be shared with the research community, IT industry, and users, which will help researchers and practitioners be aware of the importance of explainability in high-stakes anomaly detection tasks. Research results will be disseminated in data mining and security conferences, journals, books, and the project website. Additionally, we plan to develop and offer tutorials on trustworthy anomaly detection. Meanwhile, the proposed research will be integrated with education. This research will involve graduate and undergraduate students through courses and thesis projects to expand their knowledge and skills in machine learning, especially explainable machine learning. The PI will encourage the participation of undergraduate and minority students in research. The PI will also reach out to the K-12 students to demonstrate the concepts of explainable machine learning.