Trustworthy anomaly detection for smart manufacturing

Li, Z.

Summary

This dissertation focuses on the development of trustworthy anomaly detection methods aimed at improving manufacturing processes, particularly in high-tech systems. While the primary focus is on smart manufacturing, many of the techniques and findings have broad applicability beyond this domain, with several approaches designed for general-purpose use. Given the complexity of achieving trustworthy anomaly detection, this work approaches the challenges from two complementary perspectives: data-centric AI and model-centric AI, each leading to a research question that consists of multiple sub-questions.

From the *data-centric AI perspective*, the dissertation approaches the challenges related to complex and high-dimensional data, both of which are prevalent in smart manufacturing environments. The first research question (Q1.1) deals with the problem of detecting and explaining anomalies in system logs. To solve this, we developed *Logs2Graphs*, a novel method that first transforms event logs into informative graphs. These graphs are then analyzed using a specialized graph neural network model, *OCDiGCN*, to identify anomalies. The method *Logs2Graphs* not only improves detection accuracy but also provides clear explanations by providing the most relevant nodes in the anomalous graph, with the potential to facilitate root cause analysis.

The second research question (Q1.2) focuses on managing high-dimensional data, which can undermine traditional log anomaly detection methods. We developed a novel approach to reduce the complexity of log data by identifying relevant log events that are highly associated with system faults. This method improves fault detection and prediction accuracy, as demonstrated in experiments with real-world datasets. By selecting relevant system log events, the model becomes more effective at identifying gradual fault patterns that might otherwise be obscured by irrelevant data.

From the *model-centric AI perspective*, this dissertation alleviates challenges related to explainability, robustness, generalizability, and automatability of anomaly detec-

tion models. The first model-centric research question (Q2.1) investigates how to achieve intrinsic explainability in anomaly detection systems. We proposed *QCAD*, a contextual anomaly detection method based on Quantile Regression Forests, which explores dependencies between features to identify and explain anomalies. This method enhances both detection accuracy and interpretability, allowing domain experts to understand why specific objects are considered anomalous.

A related research question (Q2.2) explores the robustness of post-hoc explanations for graph neural networks (GNNs) under adversarial attacks. We found that current post-hoc GNN explanation methods are highly vulnerable to small perturbations in graph structures, which can drastically alter explanations without changing model predictions. To achieve this, we proposed *GXAttack*, one of the first optimization-based adversarial attack methods targeting GNN explanations. This research exposes the fragility of widely used GNN explainers, highlighting the need for more robust interpretability methods in high-stakes applications.

The third model-centric research question (Q2.3) deals with generalizing anomaly detection models to new, unseen environments. We developed *ARMET*, an unsupervised domain adaptation method that uses labeled normal graphs from a source domain to detect anomalies in an unlabeled target domain. By learning appropriate graph representations through adversarial learning, ARMET achieves superior performance in cross-domain anomaly detection, making it particularly useful for evolving systems, such as those undergoing software updates or hardware modifications.

The final research question (Q2.4) focuses on automating the tuning of hyperparameters in unsupervised anomaly detection systems, where labeled data is scarce. Many self-supervised learning (SSL)-based approaches to anomaly detection are sensitive to hyperparameter settings, leading to overestimated performance when labels are improperly used for tuning. To address this, we introduced *AutoGAD*, the first automated hyperparameter selection method for SSL-based graph anomaly detection. AutoGAD uses an internal evaluation metric to select hyperparameters without relying on labels, thereby mitigating label leakage and improving model reliability.

In summary, this dissertation presents a comprehensive framework for improving anomaly detection in smart manufacturing by integrating data-centric and model-centric AI approaches. We demonstrate that anomaly detection in smart manufacturing can be enhanced by using graph neural networks to handle complex log data and employing feature selection to manage high-dimensionality. Moreover, explainable models like *QCAD* help make anomaly detection more interpretable, while *GXAttack* investigates the robustness of post-hoc GNN explanations and *ARMET* enables

adaptability across domains. Lastly, automated hyperparameter tuning via *AutoGAD* supports the development of reliable anomaly detection systems without relying on labels. These contributions not only improve the reliability and efficiency of manufacturing processes but also provide insights applicable to a wide range of fields that rely on anomaly detection.