Ph.D. in Information Technology
Thesis Defense

March 20th, 2025
at 10.00 am
Sala Conferenze Emilio Gatti – building 20

Filippo LEVENI – XXXV Cycle

Structure-based Anomaly Detection and Clustering

Supervisor: Prof. Giacomo Boracchi

Abstract:

Anomaly detection is a challenging problem encountered across various application domains, including healthcare, manufacturing, and cybersecurity. Numerous statistical methods have been proposed in the literature, each relying on specific assumptions about the data being analyzed. Many of these approaches are unsupervised, where anomalies are identified as samples that deviate from the expected patterns based on the underlying assumptions. This thesis presents new unsupervised solutions for the anomaly detection problem in two different settings. In the first part of the thesis, we address anomaly detection from the general perspective of identifying samples that do not conform to structured patterns. In this case, we focus on the scenario where genuine data can be described by low-dimensional manifolds, while anomalous data cannot and are therefore unstructured. Our main contribution to the research on structure-based anomaly detection is Preference Isolation Forest (PIF), a novel anomaly detection framework that combines the advantages of adaptive isolation-based methods with the flexibility of Preference Embedding. The main intuition is to embed the data into a high-dimensional Preference Space, by fitting a collection of low-dimensional manifolds, and to identify anomalies as the most isolated points within this space. We propose two different approaches to identify isolated points: Voronoi-IForest, which leverages on suitable distances between preferences to build an ensemble of isolation trees, and RuzHash-IForest, which exploits a Locality Sensitive Hashing (LSH) scheme to avoid the explicit computation of distances, hence reducing the computational complexity of the framework. Furthermore, we propose a sliding window-based extension of PIF, namely Sliding-PIF, which leverages a locality prior to address situations where the global nature of the low-dimensional manifolds is unknown, enabling anomaly detection when only local information is available. Experiments on both synthetic and real-world datasets demonstrate that PIF successfully discriminates between structured and unstructured data, and favorably compares with state-of-the-art anomaly detection techniques. We extend our work to a scenario closely related to structure-based anomaly detection, namely structure-based clustering. In structure-based clustering, the focus is on recovering individual genuine structures rather than identifying anomalies, which are usually detected as a byproduct of the structures recovery process. In the literature, structure recovery is typically addressed for the single-family scenario, where individual genuine structures can be described by a specific model family, whereas the multi-family structure recovery scenario has been much less investigated. We propose MultiLink, a novel structure-based clustering algorithm that simultaneously deals with multiple families of models in datasets

contaminated by noise and outliers. In particular, MultiLink considers geometric structures defined by a mixture of underlying parametric models, and tackle the structure recovery problem by means of preference analysis and clustering. MultiLink combines on-the-fly model fitting and model selection in a novel linkage scheme that determines whether two clusters are to be merged. The resulting method features many practical advantages over traditional preference based methods, being faster, less sensitive to the inlier threshold, and able to compensate limitations deriving from initial models sampling. Experiments on several public datasets show that MultiLink performs competitively with state-of-the-art alternatives, both in single-family and multi-family problems. In the second part of the thesis, we focus on the more traditional setting of detecting anomalies as samples that lie in low-density regions, namely density-based anomaly detection. We start by addressing the challenging scenario where data comes as an endless stream that may exhibit dynamic behavior, and anomalies must therefore be detected in an online way. Anomaly detection literature is abundant with offline methods, which require repeated access to data in memory, and impose impractical assumptions when applied to a streaming context. Existing online anomaly detection methods generally fail to address these constraints, resorting to periodic retraining to adapt to the online context. We propose Online-IForest, a novel method explicitly designed for streaming conditions that seamlessly tracks the data generating process as it evolves over time. Online-IForest models the data distribution using an ensemble of multi-resolution histograms that track point counts within their bins, incorporating a dynamic mechanism to adjust histograms resolution and incrementally adapt to the data generating process. Experimental validation on real-world datasets demonstrate that Online-IForest is on par with online alternatives and closely rivals state-of-the-art offline anomaly detection techniques that undergo periodic retraining. Notably, Online-IForest consistently outperforms all competitors in terms of efficiency, making it a promising solution for applications where fast identification of anomalies is of primary importance such as cybersecurity, fraud and fault detection. Finally, we address a relevant anomaly detection problem in the industrial scenario, specifically the cybersecurity challenge of identifying new malware families. Classifying a malware into its respective family is essential for building effective defence against cyber threats, enabling cybersecurity organizations to detect, respond to, and mitigate malicious activities more efficiently. However, the steady emergence of new malware families makes it difficult to acquire a comprehensive training set that encompasses all classes needed for training machine learning models. Therefore, a robust malware classification system should accurately categorize known classes while also being able to detect new ones, an anomaly detection challenge referred to in the literature as open-set recognition. We propose, for the first time, to combine a tree-based Gradient Boosting classifier, which is effective in classifying high-dimensional data extracted from Android manifest file permissions, to an open-set recognition technique developed within the computer vision community, namely MaxLogit. Our approach can be seamlessly applied to a classification pipeline based on boosted decision trees, without even affecting the classification workflow. Experiments on public and private real-world datasets demonstrate the potential of our method, which has been deployed in Cleafy’s business environment and is currently part of their engine.

PhD Committee

Prof. Vincenzo Caglioti**, Politecnico di Milano**

Prof. Andrea Fusiello, **Università degli Studi di Udine**

Dr. Cristiano Cervellera, **CNR Genova**