Machine Learning CS 460 - Project

ANOMALY DETECTION USING MACHINE LEARNING ALGORITHMS

by Samir Dileep, under guidance of Dr. Subhankar Mishra
School of Computer Sciences, NISER Bhubaneswar

samir.dileep@niser.ac.in

Introduction

Anomaly detection, a crucial area within the realm of machine learning and data analytics, is focused on the recognition of patterns in data that exhibit substantial deviation from the standard or anticipated behavior. As organizations increasingly depend on extensive amounts of data for their decision-making processes, the capacity to efficiently identify anomalies becomes critical for upholding data integrity, ensuring security, and improving operational effectiveness. The realm of anomaly detection techniques encompasses a wide range of algorithms and methodologies, each designed for specific data types, contexts, and uses.

This academic poster presents a comprehensive analysis of anomaly detection methodologies, highlighting the prominent algorithms employed in sectors such as finance, cybersecurity, healthcare, and industrial systems.

Classification of Anomalies

The nature, amount and dependencies of anomalies in a given dataset are hugely dependent on the nature of said dataset. Nevertheless, we can broadly classify anomalies into three types, on the basis of the scale of the anomaly.

I. Point Anomaly: Point anomalies are defined as those instances in data which are considerably dissimilar from all points in the entire dataset. It is also referred to as global anomalies, and are the simplest and most studied so far.

II. Contextual Anomaly: Contextual anomalies differ from point anomalies in that they may be similar to non-anomalous data instances in all but a few contexts, *i.e.*, they may differ from the acceptable data in the values for a few feature dimensions. There are two attributes for these anomalies: anomalous attributes determine the context in which the anomaly is exposed, and behavioural attributes quantify the non-contextual feature values. These are usually seen in spatial and time-series data.

III. Collective Anomaly: This is a collection of instances which differ from the entire data. The instances in the collection may be related by one or more similar feature values. These are commonly observed in sequence, graph and spatial data.

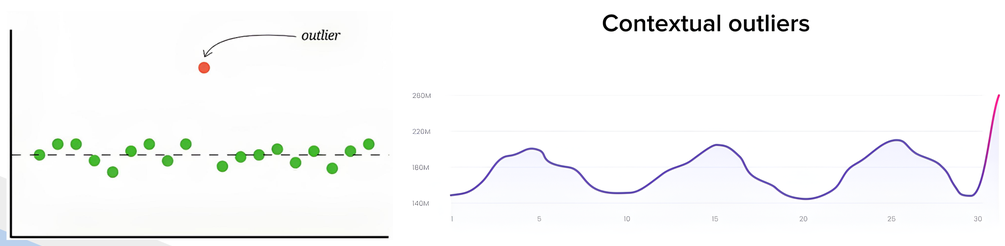

FIGURE 1 Point and Contextual anomalies in data.

Machine Learning Algorithms

Most algorithms use a similar approach to detect anomalous instances: All instances' feature values are compared to a given distribution (usually Gaussian) and those outside the distribution are considered anomalous. Alternatively, given labelled samples, an optimal distribution can be generated.

Availability of Labelled Data

I. Supervised: Here, all instances should be labelled as normal ones or anomalies. The approach is to build a predictive model for both anomaly and normal classes and then compare these two models to build a classifier.

II. Unsupervised: Here, the instances need not be labelled at all. The model assumes that outliers would be farther away from its nearest neighbours compared to normal cases. Either a fixed distribution is used to remove outliers, or they are removed based on a case-dependent distance or density metric.

There are also semi-supervised and self-supervised algorithms which are gaining popularity, due to them not needing fully labelled data.

Techinques Employed

I. Classification-based: Decision tree, SVM (single or multi-class), GMM, tree-based GP used as classifier.

II. Nearest-neighbour-based: Anomaly score calculated using distance-based (KNN) or density-based (LOF) metrics.

III. Clustering-based: Clusters containing normal data made using distance (K-means), density (HDBSCAN, PIDC), hierarchical (MST, BIRCH, CURE, CHAMELEON) or grid-based (STING, WaveCluster, DClust, GBOD) algorithms.

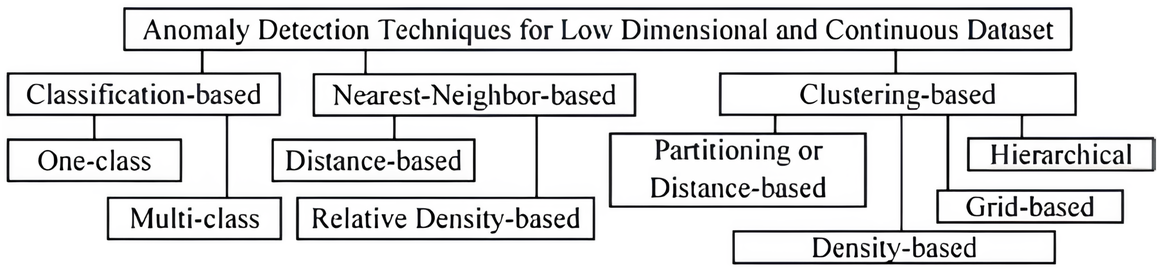

FIGURE 2 General algorithms seen to be used for anomaly detection in literature. Here, algorithms are shown for low-dimensional data, algorithms like PCA should be run on high-dimensional datasets.

Limitations

- For supervised anomaly detection, generating datasets for both the normal and anomaly class have proved to be very difficult.
- Unsupervised algorithms assume normal cases are much more common than outliers. If this is not the case, huge number of false alarms could be generated.
- Varying and generating new distribution on obtained new data turns out to be very slow.
-

References

[1] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab. Machine learning for anomaly detection: A systematic review. *IEEE Access*, 9:78658–78700, 2021.

[2] A. R. Yeruva, P. Chaturvedi, A. L. N. Rao, S. C. DimriL, C. Shekar, and B. Yirga. Anomaly detection system using ml classification algorithm for network security. In *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, pages 1416–1422, 2022.

[3] M. Zhao and J. Chen. A review of methods for detecting point anomalies on numerical dataset. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 1, pages 559–565, 2020.