Towards Anomaly Detection in the Presence of Data Drift

Jongjun Park

Computing and Software
McMaster University
parkj182@mcmaster.ca

1 Introduction

Data inevitably changes to reflect user activity and preferences, and changes in the environment. Identifying the inherent patterns to understand how data changes is a fundamental task in time series analysis and prediction. The rate and interval at which data changes, and the duration of the change, may or may not, be expected. When an input data distribution changes, this is often referred to as concept drift. Existing time series, anomaly detection techniques ignore concept drift, assuming time series concepts are stationary, and that data values follow a fixed probability distribution [1]. However, this assumption does not hold in practice. For example, temperature changes between seasons demonstrate a gradual increase from winter to spring, changes in workplace electricity usage from weekday to weekend exhibit an abrupt decrease due to a change in employee work patterns, and a company's stock price changes due to political and economic events, and investor sentiment and speculation.

Anomaly detection methods focused on (1) learning normal behaviour by searching for repeated patterns, and (2) predicting future behaviour based on historical data using deep neural networks. Both techniques misclassify data drift as anomalies leading to an increased number of false positives. On the other end, existing drift detection methods assume a negligible amount of anomalies, or fail to consider them at all. Recent work have tried to differentiate anomalies and change points [3], but still assume: (1) anomalies are short-lived and independent; (2) do not consider broader definitions of data drift; nor (3) varying baselines of normal behaviour.

2 Research Challenges

Temporal normality and localized anomalies. Previous methods define normal patterns independent of time, i.e., an observation that is similar to a normal pattern will be classified as normal despite having no similar occurrences in its immediate time vicinity. We argue that local context is extremely important in anomaly detection, such that normality must consider temporal context.

Multiple normality. Time series may exhibit multiple normal patterns within a certain time interval, with distinct patterns. Previous methods compute the anomaly score as a weighted average of distances from each normal patterns [1, 2]. However, when normal models are sufficiently different, and corresponding weights are imbalanced, subsequences that are similar to *minor* normal patterns can be misclassified as anomalies.

3 Proposal and Contributions

To address the challenges outlined, we propose an online anomaly detection method based on multiple normal patterns. First, we define the normal patterns from subsequences that appear consecutively in the training set. To capture the temporal properties of each pattern, we set correlated parameters, including a similarity threshold (based on distribution differences) and frequency within a moving window W. When a subsequence S_j arrives, its anomaly score is calculated by comparing it to the normal patterns. Simultaneously, the distances computed from each pattern are used to update the membership of S_j, adjusting the activeness of each pattern within the current time window. If anomaly scores keep high over the window, the method checks for repeated patterns and identifies new normal patterns when necessary.

References

[1] Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. Unsupervised and scalable subsequence anomaly detection in large data series. *The VLDB Journal*, 30(6):909–931, nov 2021.

[2] Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J Franklin. Sand: streaming subsequence anomaly detection. *Proceedings of the VLDB Endowment*, 14(10):1717–1729, 2021.

[3] Kim-Hung Le and Paolo Papotti. User-driven error detection for time series with events. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 745–757, 2020.