TRADE: Anomaly Detection in Sparse and Irregular Time Series

Janne Cools [1], Toon Boeckling [1], and Antoon Bronselaer [1]

[1] Ghent University, Belgium
{janne.cools,toon.boeckling,antoon.bronselaer}@ugent.be

Finding anomalies in time series is an important research topic [2]. Most state-of-the-art anomaly detection algorithms rely on seasonality and temporal locality. However, these properties are hard to exploit in *sparse* and *irregular* time series, such as electronic health records or Wikipedia page histories, where observations typically are rare and unevenly spaced.

As a solution, we introduce TRADE (Transition-Rule-based Anomaly Detection Engine), a framework for detecting point anomalies (i.e., single erroneous values) using logical constraints. TRADE extends the *selection rule framework* [1], which expresses forbidden value combinations within a single record by means of selection rules, with transition rules that express constraints over two consecutive records (e.g., "a person's age may not decrease"). Anomalies are identified through minimum-cost repair: the attribute value modifications that lead to satisfaction of all rules at the lowest total cost, with customizable costs per attribute, are marked as anomalous. Since finding an optimal repair is NP-hard, TRADE employs a greedy strategy. After a one-time, data-independent preprocessing step, it starts at an *anchor* record and scans forward and backward through the series, repairing one record at a time given its already-consistent neighbor. Starting from multiple anchors increases the likelihood of finding a global minimum-cost repair, and an early-stop mechanism reduces their cost.

We evaluate TRADE on two synthetic datasets (NBA player statistics and gait-analysis metadata) and two real-world datasets, scraped from Wikipedia infobox histories (elections and settlements). We compare it against two time-series-specific and two general-purpose error detection baselines. TRADE achieves the best F_1-scores in most settings, for example 0.74 for elections compared to its best baseline of 0.41. Restricting TRADE to selection rules lowers its recall metric (e.g., from 0.63 to 0.43 for NBA), demonstrating the added value of using selection rules. Preprocessing is a one-time cost that depends on the rule set, while detection time scales linearly with the dataset size, with efficiency optimizations due to the early-stop mechanism (going from 14% to 44%). Limitations include the detection of column-wide errors (i.e., an attribute that has a wrong value at every timestamp) and the inability to express allowed changes that depend on the time gap between records, since rules currently use fixed constants. We plan to address the latter with variogram-based rules in future work.

References

[1] Toon Boeckling, Guy De Tré, and Antoon Bronselaer. Cleaning data with selection rules. *IEEE Access*, 10:125212–125229, 2022.

[2] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: a comprehensive evaluation. In *Proc. VLDB Endow.*, volume 15, page 1779–1797. VLDB Endowment, 2022.