Comparative Evaluation of Unsupervised Machine Learning Algorithms for Anomaly Detection in Time Series Data

H. Asela*

Department of Electrical and Information Engineering, University of Ruhuna, Galle, Sri Lanka
**Corresponding Author E-mail: asela.h@eie.ruh.ac.lk, TP: +94716637905*

Anomaly detection is a mechanism of identifying data, occurrences and observations deviating from the normal pattern. Anomaly detection in time series data is used to identify critical and fraudulent events, technical glitches and potential opportunities in the systems. Hence it is important to build robust models that can properly identify anomalies in time series data. In the literature, Anomaly detection is done using supervised, unsupervised and hybrid machine learning algorithms. However, most researches have focused on unsupervised algorithms to build anomaly detection models due to unavailability of labelled data. These unsupervised algorithms are based on probability, distance, density or a boundary function. This study provides a comparative evaluation of mutltiple unsupervised algorithms for anomaly detection in time series data, namely Elliptic Envelope, Gaussian Mixture Model, Isolation Forest, Local Outlier Factor, One Class Support Vector Machine and K-Means Clustering algorithm. Based on previous literature, these algortihms were selected as a famously used subset of algorithms for multi-domain anomaly detection. The algorithms were evaluated using Yahoo! Webscope S5 labeled dataset. This dataset contains real and synthetic time series data in 4 classes with overall 572,966 data instances and 367 metrics. Feature extraction was done using time series decomposition and statistical techniques. These extracted features were integrated with specific features given in the data classes to improve the performance of these algorithms. The feature normalization was done using min-max scale. Elliptic Envelope and Gaussian Mixture Model were the best performing algorithms with 26.3% - 81.7% F1 score, 26.4% - 82.7% true positive rate and below 2% false alarm rate for the 4 data classes in the dataset. The reason for this is the ability of probabilistic models to adapt and identify the complex patterns in time series data that helps to identify deviations in a more robust way. One Class Support Vector Machine is the worst performing algorithm with 1.2% - 6.5% F1 score and around 50% false alarm rate for the data classes in the dataset as its decision function was unable to properly adapt to the complex patterns in time series data. However, it had 96.2% - 99.5% true positive rate. Other algorithms performed moderately where Isolation Forest performed best in the high contamination data class

***Keywords*:** Anomaly detection; Time series data; Unsupervised machine learning algorithms; Comparative evaluation