Anomaly Detection in Time Series Dataset

Atul Kumar Jha, Master of Science in Computer Science

University of Dublin, Trinity College, 2020

Supervisor: John Waldron

In terms of data mining, anomalies are defined as a data object that vary significantly from the rest of the values, as if it was generated by a different mechanism. Early detection of anomalies plays a vital role in any organisation. To maintain the consistency of an individual's data and to protect any corporation against malicious attacks, it is important to detect anomalies as early as possible. Due to security and privacy issues there are very limited real-world dataset available that can be used to benchmark the anomaly detection models based on their test score. The available datasets are the both real and synthetic some of them being highly imbalanced. Here the paper purposes different kind of approach which can be used when dealing with the imbalanced dataset while maintaining the integrity of the dataset as well as to differentiate which kind of algorithms works best on different types of dataset.

In this study, various methods have been performed that can be used when working on anomaly detection, as well as comparing their efficiency. The proposed system uses different algorithms like Linear Regression, Ridge Regression, ARIMA model, Long-Short-Term-Memory, Gaussian Distribution method, Auto regression model, etc. For evaluation purpose different metrics such as confusion matrix, ROC curve, precision, recall were used. The result showed that K-means model score was able to maintain good score on both Yahoo S5 and Numenta dataset, However ARIMA worked well only on Yahoo S5 dataset. For the Credit Card dataset, the most difficult part was to balance the dataset once balanced, the result showed that Logistic regression worked well on it whereas other algorithms struggled to classify correctly.