Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

R&D Project Proposal

ANOMALY DETECTION IN TIME SERIES DATA

Kaushik Manjunatha

Supervised by

Prof. Dr. Paul G. Plöger
Dr. Anastassia Küstenmacher

December 2020

1 Introduction

Anomaly detection has been a research topic for a long time. In this world of digitization, the amount of data transferred exceeds the ability of humans to study it manually. So, data analysis will become a necessity. One of the most important data analysis tasks is the detection of anomalies in data. Anomalies are data points that deviate from the normal distribution of the whole dataset, and anomaly detection is the technique to find them. In a dataset of network activities, an anomaly can imply an intrusion attack[1]. Anomalies are not limited to detection of intrusion attacks, an anomaly in a financial transaction hints a financial fraud, and that in a medical image hints a disease. Some other applications of anomaly detection are in detecting industrial damages, preventing data leak, identifying security vulnerabilities, military surveillance etc. Anomaly detection methods are specific to the type of data. For instance, the algorithms used to detect anomalies in images are different from the approaches used on data streams. In this project an attempt is made to identify methods for anomaly detection in time-series data.

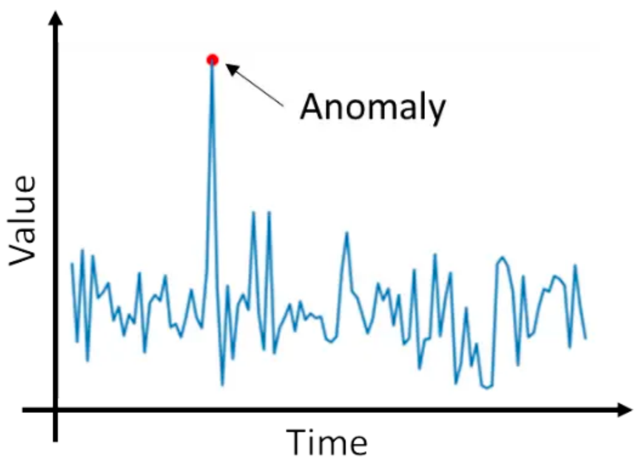

Figure 1: *Anomaly plot of time series data*

A time series is a sequence of historical measurements of an observable variable at equal time intervals [2]. Time series is an ordered collection of data points that are recorded at equal time intervals discretely or continuously. The majority of the data present in the real world contains a temporal component that depends on or varies with time. For example, weather, sound waves, a person's heart rate, etc. These time-series data are hard to interpret due to their varied features such as their patterns, internal noises, larger dimensions, etc. Time-series can be used for predicting the future based on past data or understanding the past and observing pattern of the data. Time series forecasting is an approach to predicting future values by using the past data values and analyzing the pattern. Based on the characteristics and the trends of the time series approaches, the forecasting techniques are chosen to predict the future time constraint values from the historical data. Time series analysis is a must to understand seasonality, cyclicality, trend, and randomness in the series.

Anomaly detection is an unsupervised pattern recognition task that can be defined under different statistical models. For a set of training samples containing no anomalies, the goal of anomaly detection is to design or learn a feature representation, that captures "normal" appearance patterns. Anomaly detection is an effective approach to dealing with problems in the area of network security [1]. Rapid development in technology has raised the need for an effective detection system using machine learning in order to detect novel and advanced intrusions. As there is an increase in the amount of data that is being transmitted everyday from one network to another, there is an increased need to identify intrusion in such large datasets effectively and periodically. Data mining and machine learning approaches could prove effective in this regard[3]. Anomaly or Outlier detection is an especially tricky problem in networks, financial transactions, and real-world data analysis because the statistical properties of anomalies are difficult to be assessed with low training dataset. [4].

Autoencoder is a neural network designed to learn an identity function in an unsupervised way to reconstruct the original input while compressing the data in the process so as to discover a more efficient and compressed

representation. Using autoencoders, high-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors[5].
This is an effective way of initializing the weights that allow deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.
The core idea is to create an autoencoder network to learn data representation of normal (non-malicious) data points. The decoder would try to regenerate the encodings by trying to minimize the loss in reconstruction. The encoding would be uniform for regular day to day data. The keynote here is that the data with anomalies will have higher reconstruction errors than a regular dataset.

1.1 Problem Statement

In this R&D project, the aim is to find, understand, and do comparative analysis for the different state-of-art techniques used for anomaly detection in time series measurement data.

1.2 Motivation

Time series data is not so representative in the community in the technique mentioned above. Usage of image data by autoencoders for anomaly detection in computer vision area is commonly known. This project goes a step further and uses time series data for anomaly detection on measurement data[6].
The goal is to identify and understand the different state-of-art techniques used to detect the anomalies in time series data. This is an attempt to do a comparative analysis of different state-of-art techniques between the classical machine learning approaches and deep learning techniques like autoencoders [7]. The autoencoders learn better due to their low dimensionality representation, and hence are expected to give a better outcome than machine learning approaches.

2 Methodology

The aim of this project is to build a robust model for time series anomaly detection for measurement data. Using FFT/ DCT techniques, the extraction of features such as sensor values, frequency, amplitude, phase and time stamp, etc., along with some statistical features such as mean and standard deviation is done on measurement data. This will be the input to the anomaly detection methods in order to determine the anomalies from the normal behavior and compute its accuracy with machine learning approaches such as One-Class SVM or LSTM. The same steps are repeated now with the selected, state of the art method to determine the anomalies.
The results are compared with the baseline results which give an overview of detected results [8], and goodness along with the performance of the state-of-art approaches which is very important for machine learning models to learn and to to provide better results.

3 Project Plan

3.1 Work Packages

The following are the work packages associated with this project, which are to be delivered as a whole package at the end of this project.

WP1 Literature search

WP2 Dataset collection

WP3 Implementation of state of art

WP4 Implementation of classical machine learning anomaly detection methods

WP5 Evaluation

WP6 Final report

	Work package	Task
WP1	Literature review	Study on different methods of anomaly detection
		Identify the method to implement anomaly detection
WP2	Dataset collection	Identify the dataset to implement the state of the art - autoencoders
		Preprocessing of the data
WP3	Implementation of state of art	Familiarising the tools and packages to be used
		Anomaly detection using state of art methodolgies by autoencoders
WP4	Implementation of classical machine learning anomaly detection methods	Extraction of features
		Anomaly detection using machine learning approach by LSTM or one class SVM
WP5	Evaluation	Evaluation of the detected anomalies in state of art method
		Evaluation of the detected anomalies in machine learning approach
		Comparison of the state of art and machine learning approach
WP6	Final Report	Determine the future work and improvements on the comparitive results
		Final report

Figure 2: *Work package*

3.2 Milestones

M1 Literature search

M2 Dataset finalization

M3 Study about tools and packages

M4 Implementation of state of the art for anomaly detection

M5 Implementation of FFT/ DCT method for feature extraction and machine learning methods for anomaly detection

M6 Comparison of performance of machine learning methods with state of the art for accuracy of the detected results

M7 Report submission

3.3 Project Schedule

The overall research work target period is provided in Figure 3.

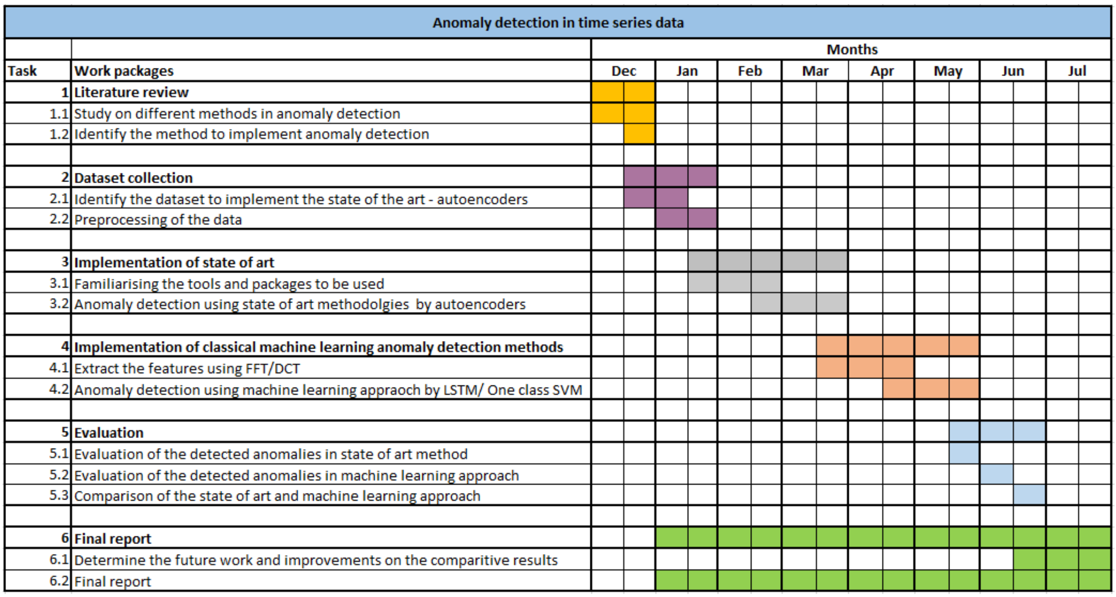

Figure 3: *Research work timeline*

3.4 Deliverables

Minimum Viable

- Literature review
- Analysis of state of the art
- Dataset collection and analysis
- Data preprocessing
- Implementation of state of the art for anomaly detection
- Final report

Expected

- Implementation of anomaly detection using a classical machine learning approach
- Comparison of machine learning approach and state of the art for anomaly detection

Desired

- Evaluation of the results with another variant of autoencoder

References

[1] M. Braei and S. Wagner, "Anomaly detection in univariate time-series: A survey on the state-of-the-art," 2020.

[2] T. Kieu, B. Yang, C. Guo, and C. S. Jensen, "Outlier detection for time series with recurrent autoencoder ensembles," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 2725–2732. [Online]. Available: https://doi.org/10.24963/ijcai.2019/378

[3] D. Li, D. Chen, J. Goh, and S. kiong Ng, "Anomaly detection with generative adversarial networks for multivariate time series," 2019.

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, Jul. 2009. [Online]. Available: https://doi.org/10.1145/1541880.1541882

[5] R. Yao, C. Liu, L. Zhang, and P. Peng, "Unsupervised anomaly detection using variational auto-encoder based feature extraction," in *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 2019, pp. 1–7.

[6] M. Landauer, M. Wurzenberger, F. Skopik, G. Settanni, and P. Filzmoser, "Time series analysis: Unsupervised anomaly detection beyond outlier detection," in *Information Security Practice and Experience*, C. Su and H. Kikuchi, Eds. Cham: Springer International Publishing, 2018, pp. 19–36.

[7] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLOS ONE*, vol. 11, no. 4, pp. 1–31, 04 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0152173

[8] N. Goix, "How to evaluate the quality of unsupervised anomaly detection algorithms?" 2016. [Online]. Available: https://arxiv.org/abs/1607.01152