Unsupervised Machine Learning Anomaly Detection for Monitoring

IKHTIAR AHMED*, Technical University of Dortmund, Germany

Earlier, the monitoring and maintaining system was more manual and static. Nowadays it has become a huge concern due to the high rate of data production. That is why monitoring systems are getting larger, dynamic and dependent on artificial intelligence. In this report, I have focused on unsupervised anomaly detection in time series data with several statistical process control (SPC) models. I have also introduced some advanced technology for anomaly detection like, ARIMA family and Engineering process control (EPC). During this study, I have found that different methods and functions can be used for a single problem solution. This study indicates which charts and methods are better for time series data including their disadvantages.

CCS Concepts: • **Anomaly Detection**; • **Statistical Process Control (SPC)**; • **Cumulative Sum Control (CUSUM)**; • **ARIMA**;

ACM Reference Format:
Ikhtiar Ahmed. 2022. Unsupervised Machine Learning Anomaly Detection for Monitoring. 1, 1 (February 2022), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Anomaly (Outliers) detection is a technique to identify the unusual behavior or pattern of the system. In other words, we can say 'Anomaly' means unexpected event or value of the dataset. For a huge number of applications, detecting anomalies is a significant issue. On a daily basis, we are using various applications for personal or business purposes; between them in many applications; the information is made by several processes which could help to produce the exact outcome of the system. Constantly, it is not possible for machines to work perfectly and produce accurate outcomes. The outcome of the system depends on the data. In the generating process, sometimes data contains noise that can provide unusual behavior or pattern which creates anomalies or outliers of the system [1]. It can detect useful information like, unusual metric that someone is observing in the system.

Statistical process control is generally utilized as a quality control technique in the industry [1]. The main goal of statistical process control (SPC) is to process quality and keep up with the interactions to the fixed target value. SPC have several control charts which are used for unusual variation of the manufacturing process and indicate the problem. Control charts normally designed for monitoring the data and it shows the characters of data. From these charts, models are generated so that we can differentiate the anomaly from a prediction. During the implementation of these methods, several output characteristics appear like, error rate, throughput, latency, concurrency and utilization.

The broad objective of this report is to discuss anomaly detection using Statistical Process Control, its advantages and disadvantages with a real life example. The specific objective of my study is to describe other useful methods for anomaly detection in time series data. Here I have found that SPC methods are being improved according to the

Author's address: Ikhtiar Ahmed, ikhtiar.ahmed@tu-dortmund.de, , Technical University of Dortmund, Poststelle: August-Schmidt-Straße 1 Rektorat:, August-Schmidt-Straße 4, Dortmund, North Rhine Westphalia, Germany, 44227.

Manuscript submitted to ACM

demand and variation of time series data. For multivariate data, the ARIMA family has the most advanced functions to generate models. At the end, I can say that day by day anomaly detection is a topic that is growing and moving faster with advanced technologies.

2 LITERATURE REVIEW

An unsupervised machine learning dataset is normally unstructured or unlabeled. They are used for several machine learning techniques to process the pattern of an unstructured dataset. Many researchers have already used different techniques to identify anomalies. Between them statistical process control (SPC) is one of the most commonly utilized methods to monitor a variety of manufacturing systems. There are some recent developments of the SPC model which includes some control charts such as Shewhart control chart, Moving control chart and cumulative sum control chart (CUSUM), are examples of techniques that can be applied in this method [8]. Though this kind of control chart only works for univariate data.

In this section, i will describe some previous literature, related to my works are described as follows – In the work of [9] Richard Stone "Time series models in statistical process control: considerations of applicability" focused on the autocorrelation problem in time series data. The author of the paper discussed some limitations of the SPC model. We know, SPC model includes some control charts that are not efficient to detect anomalies from the autocorrelation time series data. The Shewhart control chart is the basic process to present the behavior of the data , but it has some control limits.

A statistical control represents the data in several control charts. SPC is frequently used as a quality control method in industry where it mostly tracks the univariate data. Research by Lee-Ing Tong et al. [10] focused on this factor. They wanted to process the multivariate data using SPC method but the traditional SPC method mislead the results. When the correlation is present in the multivariate data then the characteristics of the increase the type I and type II error. But, there are some multivariate control charts to solve this problem such as Hotelling's T2 control chart, multivariate cumulative sum, but practically they also have some drawbacks [10].

From this, to solve this kind of problem I started to research and found some possible solutions. To solve the Autocorrelation problem in time series data you can use ARIMA (Autoregressive integrated moving average) family and also for multiple variate data Engineering process control helps to detect the anomalies efficiently. This engineering process control (EPC) technique reduces the process variability and tries to keep the process output near to the desired level [10].

3 METHODOLOGY

To detect anomalies in a system, prediction and modeling is the primary step to follow. Prediction is an idea that comes from an existing model. Model sets a standard limit to be compared with the data set. Identifiers, unusual cases can be predicted and detected with the help of models. Models can be found in different ways. To work on time series data, Statistical Process Control is highly used to generate models as it is ubiquitous. Statistical Process Control is widely used for detecting abnormal behavior, monitoring processes, and improving product quality [6]. Control charts are very common in SPC methods. Different charts are developed which are used in different sectors like business, economics, healthcare, environment and so on. The Moving window control chart changes the mean after each value change. Disadvantages of using the moving window control chart is that we need to track recent data history to fit them into the window. That is why the EWMA chart is more suitable for time series data. The EWMA chart can detect smaller shifts and it uses the weighted average of every single observation of any time [6]. CUSUM chart is kind of similar to EWMA but it is a slower and less appropriate monitoring scheme. On the other hand, CUSUM is better in mean shift detection.

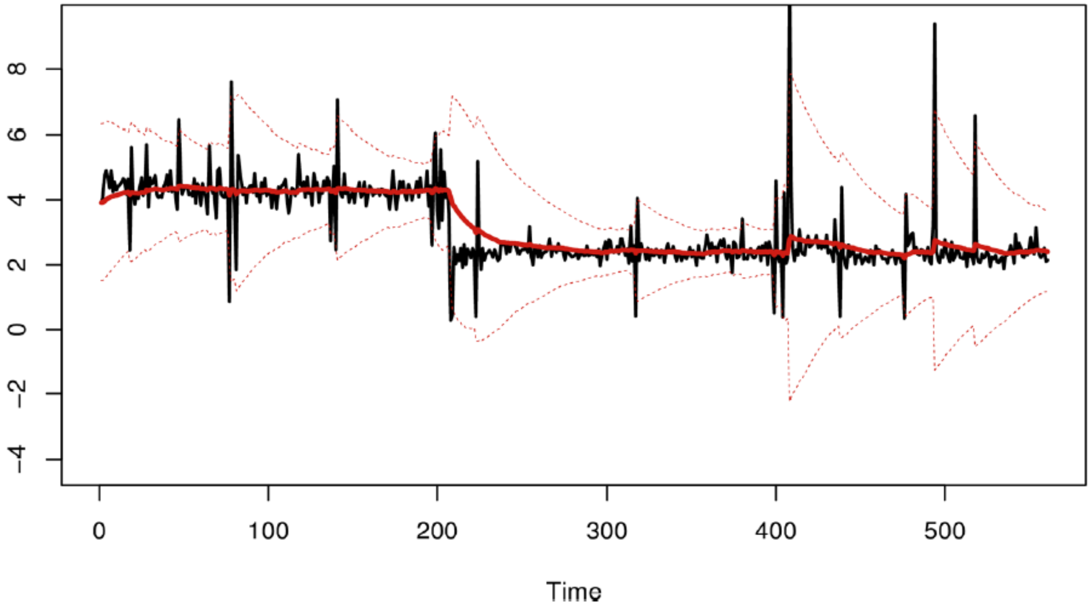

Fig. 1. Exponentially Weighted Moving Average

EWMA model is mainly used to describe a time series. In the control schemes of EWMA, the moving average is resulted from the multiplying historical observations where the weight decays exponentially over time [7].

$$EWMA_t = \alpha * r_t + (1 - \alpha) * EWMA_{t-1}$$

Here,

Alpha = The weight which is user-defined

r = Indicates the value of the series in current period

EWMA provides observations depending on previous observations, thus it is a recursive function [4]. Unlike the moving window control chart, it does not require remembering historical data. It is memory and CPU efficient. Points can be detected outside the range and values are generated with the mean by the Gaussian distribution process. Due to some disadvantages of the EWMA chart, another advanced model has arrived named ‘ARIMA’ [6].

In the time series metrics, trends and seasonality are two characteristics that break many models [6]. A continuous increment or decrement in the curve is called trend and it does not repeat. Seasonality is a pattern that repeats in a system [11]. There are different approaches to understanding trends. While trend often refers to historical changes of data, trend is nothing that happened in the past (this is more like a historical drift), but trend implies a prediction of future behavior. Or, in other words, a positive trend means that it is likely that the growth continues. The presence of variations that occur at regular intervals is very common in real-world datasets and identifying these patterns help improve our anomaly detection and forecasting efforts.

Trends and seasonality may cause false alarms. It may not determine unusual sudden events. To overcome this problem, Fourier Transform function can be used. It decomposes signals into different frequencies. For sound processing, decomposing, manipulation and combining frequencies, Fourier transform can be used [6].

In statistical time series, the ARIMA (Autoregressive integrated moving average) family of time series models and Box-Jenkins approach are very well known. It gives more accurate predictions depending on the very recent data. That is why this method is highly used in weather forecasting. We can apply parameters, various extensions and change it according to our data and desired output [6]. ARIMA models contain a lot of parameters such as p, q and d, hence it is a parametric method. It works like linear regression model where the predictors are not correlated and independent to each other [5].

Where,

- p : Number of lag observarions in the model.
- q : Indicates the size of the moving average window.
- d : Degree of differenciation.

After completing the first stage: modeling, then comes prediction. During prediction, one can find and detect the anomaly. So it is one of the foundation steps in anomaly detection.

One-step-ahead prediction is the simplest prediction type. It means that the upcoming value will be similar to the previous one. In the next level, prediction depends on the recent central tendency. In this case, the predicting value is close to the recent values. To predict from a high range of data set, simple mean and standard deviation or EWMA chart is used. To get more sophisticated results ARIMA models are a good choice. Evaluating these predictions gives a clear idea about anomalies in a system.

To summarize the whole methodology, it can be said that anomaly detection depends on proper data modeling and predicting abnormal behavior compared to the model [6]. These models are derived from historical data, which is input. Then the outputs are considered as parameters to predict upcoming events. That is how anomalies are detected in an unsupervised way with time series data.

4 IMPLEMENTATION AND DISCUSSION

In this chapter, I will discuss the behavior of the anomaly detection approaches. To apply the anomaly detection practically you need to follow two options.The first choice is to produce alerts while the other option is to record the occurrence for further study [6]. It's risky to generate an alert based on metrics because alerting anomalies will certainly generate a lot of noise. The best option is to record these unusual observations.

Sometimes, it is critical to ensure the problem that you are attempting to detect as a reliable signal. For this reason, you must identify metrics which will give you a good result when the system is normal and also provide abnormal behavior when the system is in trouble. Choosing a metric you need to check some characteristics which will give you actual output. They are -

- Error Rate.
- Throughput.
- Latency - Which is always give you a complex model distribution.
- Concurrency and utilization.

All the features are difficult to examine using normal statistical techniques without a good model. Now, you can use any sophisticated method to analyze the abnormal behavior which is best for your problem. To summarize the process, I am showing you a flowchart which helps you to make the decision to detect anomalies in a more precise way.

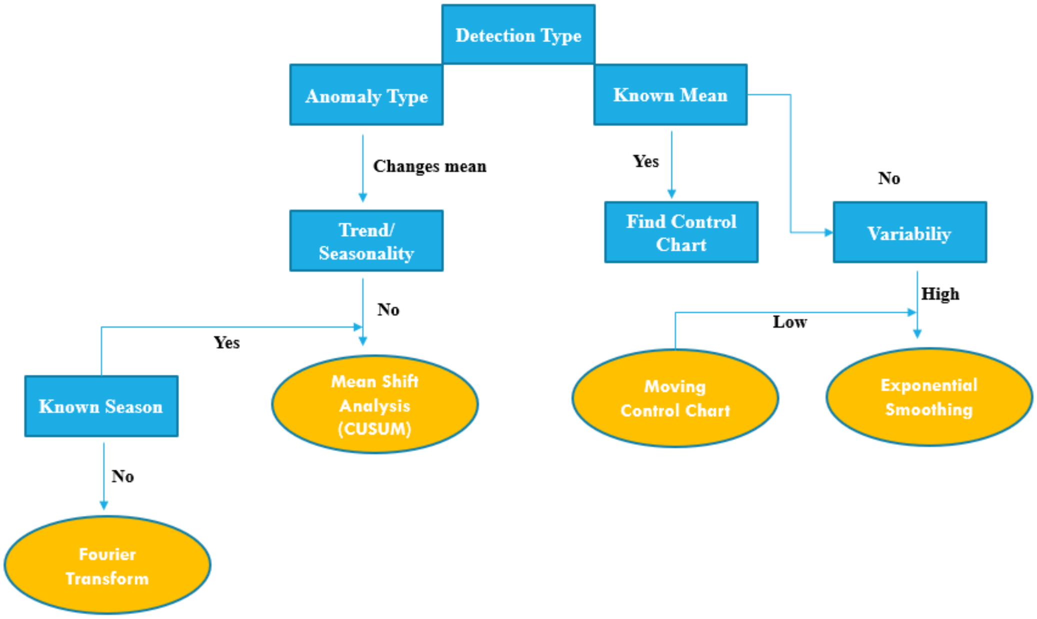

Fig. 2. A work example

In practical work, there are some limitations in every approach to detect anomalies. It is not easy to detect anomalies from time series data because it has some characteristics such as trend and seasonality. In time series data, trend always plays a vital role because it occurs in historical changes (Increasing or decreasing) over the time period [8]. To detect anomalies using statistical process control Shewhart control chart is the elementary knowledge [2]. Though there is no better substitute to detect anomalies from time series data, you need to improve some processes. From my point of view, I will suggest ARIMA (Autoregressive integrated moving average) families which include the box-jenkins approach that will provide a good prediction to detect anomalies. In this work, i only discussed anomaly detection for univariate data. But on the other hand, what will be the solution of multivariate data? The form of multivariate autocorrelation is difficult to detect and evaluate.To address this issue, multivariate CUSUM and Maltivariate EWMA control charts gives you a good result [3].

5 CONCLUSION

In this work, for the early detection of unusual behavior from time series data, I have presented a statistical method called SPC which includes various control charts. In the very beginning, I have tried to give an overview about statistical process control. Secondly, I have focused on the most fundamental part of the system which was modeling and prediction, and also have provided some knowledge which method is best to detect anomalies. Finally, I have presented a flowchart which helps give an idea on how to detect anomalies. This current method focused on developing the approach to detect anomalies from time series data for the certain application domain. Anomalies are common occurrences in today's systems. As there are different types of anomalies, as a result there are several methods to detect anomalies . This makes anomaly detection more complicated. Several methods can be conjugated to detect anomalies.

REFERENCES

[1] Charu C Aggarwal. 2016. Outlier analysis second edition.

[2] Layth C Alwan and Harry V Roberts. 1988. Time-series modeling for statistical process control. *Journal of business & economic statistics* 6, 1 (1988), 87–95.

[3] Sotiris Bersimis, Stelios Psarakis, and John Panaretos. 2007. Multivariate statistical process control charts: an overview. *Quality and Reliability engineering international* 23, 5 (2007), 517–543.

[4] corporatefinanceinstitute.com. 2015. EWMA Formula. Retrieved February 05, 2022 from https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/exponentially-weighted-moving-average-ewma/?fbclid=IwAR0Jjw2nC8Cqg5fjbU4PY_DkoUTA-7SCAIz4dLhRq73Rew_dNyI2wVDaz3c

[5] investopedia.com. 2012. ARIMA Model. Retrieved February 05, 2022 from https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

[6] Preetam Jinka and Baron Schwartz. 2016. *Anomaly Detection for Monitoring.* O'Reilly Media, Incorporated.

[7] Farid Kadri, Fouzi Harrou, Sondès Chaabane, Ying Sun, and Christian Tahon. 2016. Seasonal ARMA-based SPC charts for anomaly detection: Application to emergency department systems. *Neurocomputing* 173 (2016), 2102–2114.

[8] oraylis.de. 2015. Seasonality. Retrieved February 05, 2022 from https://www.oraylis.de/blog/2015/trend-in-times-series-analysis

[9] Richard Stone and Mark Taylor. 1995. Time series models in statistical process control: considerations of applicability. *Journal of the Royal Statistical Society: Series D (The Statistician)* 44, 2 (1995), 227–234.

[10] Leeing Tong, C Yang, C Huang, and C Shou. 2005. Integrating SPC and EPC for multivariate autocorrelated process. In *Processing of the Fifth International conference on Electronic Business, Hongkong.* 692–696.

[11] towardsdatascience.com. 2018. Trend and Seasonality. Retrieved February 04, 2022 from https://towardsdatascience.com/trend-seasonality-moving-average-auto-regressive-model-my-journey-to-time-series-data-with-edc4c0c8284b