A WAY OF ANOMALY DETECTION IN ENGINEERING EQUIPMENT CHARACTERISTICS OF SYMMETRA AT IHEP IT CENTER

E. Popova [a], V. Kotliar [b]

Institute for High Energy Physics named by A.A. Logunov of National Research Center "Kurchatov Institute", Nauki Square 1, Protvino, Moscow region, Russia, 142281

E-mail: [a] Ekaterina.Popova@ihep.ru, [b] Viktor.Kotliar@ihep.ru

The information flow should be monitored on anomaly detection. It is important, because it allows you to see a possible problem in advance and prevent it from turning into a real one. A huge flow of diverse data within the modern computing center flows from everywhere. As a rule, these are time series – numerical characteristics that are consistently measured after some time intervals.

At this work there was developed the way of analysis for engineering equipment characteristics in centralized system of uninterrupted power supply (Symmetra) at IHEP IT center. When tracking time series, extracted from the data processing and storage system, anomalies are detected using the Twitter AnomalyDetection package. The information on problem is provided to the engineering and operational staff.

Keywords: Data processing, outlier, statistics, anomaly detection, monitoring, UPS Symmetra

1. Introduction

Anomaly or outlier is an object that differs from most other objects. The percentage of anomalies in data stream may be small, but we should track them, otherwise it can lead to serious consequences. Main challenges when we try to find anomalous events or objects are:

- the boundary between normal and anomalous behavior is often not precise;
- defining normal behavior is not trivial;
- normal behavior is changeable in some knowledge areas and it requires to permanent tracing;
- anomaly is different for different application domains;
- degree to which class labels (anomaly or normal) are available for at least some of the data;
- data contains noise.

There is a broad spectrum of anomaly detection techniques. They are: classification based, clustering based, nearest neighbor based, statistical, information theoretic, spectral. The use of this or that technique depends on the type of task for anomaly detection. To formulate the task we need to determine nature of the input data, the availability or unavailability of labels for data, the constraints and requirements for knowledge area. Input data can be sequential, spacial or graph. Classical example of a sequential data is a time series. It is also important to understand the essence of the desired anomaly.

There are three types of anomalies: point, contextual, collective.

Based on the extent to which the labels are available, anomaly detection techniques can operate in one of the following three modes:

- supervised anomaly detection;
- unsupervised anomaly detection;
- semi-supervised anomaly detection.

Typically, the outputs produced by anomaly detection techniques can be scores or labels.

2. Statistical techniques

Statistical anomaly detection techniques are based on the following key assumption: “Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model” [1].

We can use both parametric as well as non-parametric techniques for model finding.

Parametric techniques assume that the normal data is generated by a parametric distribution with parameters Θ and probability density function f (x, Θ), where x is an observation. The anomaly score of a test instance (or observation) x is the inverse of the probability density function, f (x, Θ). The parameters Θ are estimated from the given data [1].

Based on the type of distribution assumed, parametric techniques can be further categorized as follows:

– *Gaussian Model* - such techniques assume that the data is generated from a Gaussian distribution. The parameters are estimated using Maximum Likelihood Estimates (MLE). The distance of a data instance to the estimated mean is the anomaly score for that instance. A threshold is applied to the anomaly scores to determine the anomalies. Different techniques in this category calculate the distance to the mean and the threshold in different ways [1]:

1. 3 σ rule - $\mu \pm 3\sigma$ region contains 99.7% of the data instances
2. The box plot rule - [Q1 − 1.5IQR; Q3 + 1.5IQR] contains 99.3% of observations (Q1 – lower quartile, Q3 – upper quartile, IQR = Q3 – Q1 – Inter Quartile Range)
3. Grubb’s test - is used to test whether or not a set of data contains an outlier.
4. The student’s t-test
5. Hotelling t^2-test
6. χ^2 statistic, etc.

– *Regression Model* anomaly detection technique consists of two steps. In the first step, a regression model is fitted on the data. In the second step, for each test instance, the residual for the test instance is used to determine the anomaly score. The residual is the part of the instance which is not explained by the regression model. The magnitude of the residual can be used as the anomaly score for the test instance, though statistical tests have been proposed to determine anomalies with certain confidence [1]. Examples:

1. Akaike Information Content (AIC)
2. Autoregressive Integrated Moving Average (ARIMA)
3. Autoregressive Moving Average (ARMA), etc.

In nonparametric techniques the model structure is not defined by default, but is instead determined from given data.

3. Twitter AnomalyDetection

For detecting anomalies in our APC Symmetra UPS system we choosed Twitter open-source R package that automatically detects anomalies in Big Data in a practical and robust way [4]. Its work is based on Seasonal Hybrid ESD (S-H-ESD) algorithm which in its turn builds upon the S-ESD algorithm (Figure 1).

The ESD test is a generalization that can test whether or not you have up to r outliers. It can answer the question, “How many outliers does the data set contain?” The principle is rather simple - it’s looking at the standard deviations of individual elements. The process is more delicate than that in Grubs’ test, because if you have two outliers, they’ll interfere with the sample mean and standard deviation, so you have to remove them after each iteration [2].

S-H-ESD uses more robust statistical techniques and metrics such as median and MAD (Median Absolute Deviation). Run time of S-H-ESD is higher than that of S-ESD and in cases where the time series under consideration is large but with a relatively low anomaly count, it is advisable to use S-ESD. [3]

```
Input:
X = A time series
n = number of observations in X
k = max anomalies (iterations in ESD)
Output:
$X_A$ = An anomaly vector wherein each element is a tuple
(timestamp, observed value)
Require:
$k \leq (n \times .49)$
1. Extract seasonal component $S_X$ using STL Variant
2. Compute median $\tilde{X}$
/* Compute residual */
3. $R_X = X - S_X - \tilde{X}$
/* Detect anomalies vector $X_A$ using ESD */
4. $X_A$ = ESD(R, k)
return $X_A$
```

Figure 1. ESD algorithm

4. From Symmetra to ES

Symmetra is a high-efficiency 3-phase UPS that is scalable as data center grows up. “Elasticsearch (ES) is an open-source, broadly-distributable, readily-scalable, enterprise-grade search engine. Accessible through an extensive and elaborate API, Elasticsearch can power extremely fast searches that support data discovery applications.“ [5] It gathers UPS metrics from two APC Symmetra PX 160kW and more than 20 APC PDU through internal feature of APC to store data on a remote ftp server. Then all these data are parsed with python programs and are put through REST API to the ElasticSearch cluster (Figure 2).

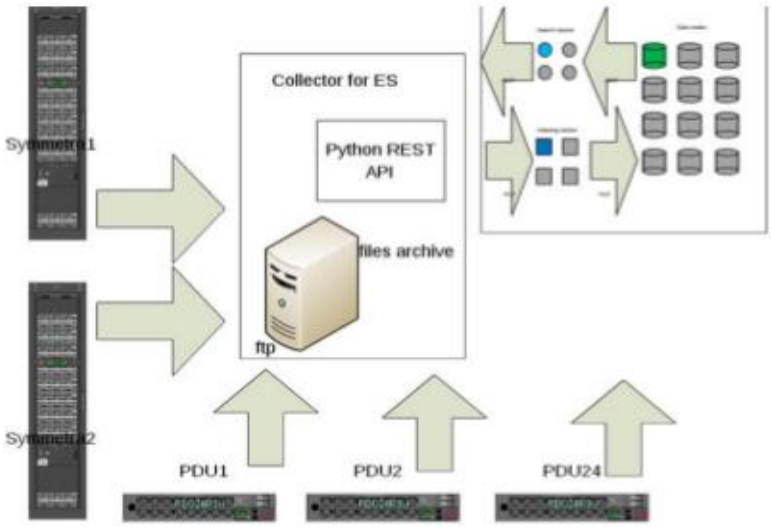
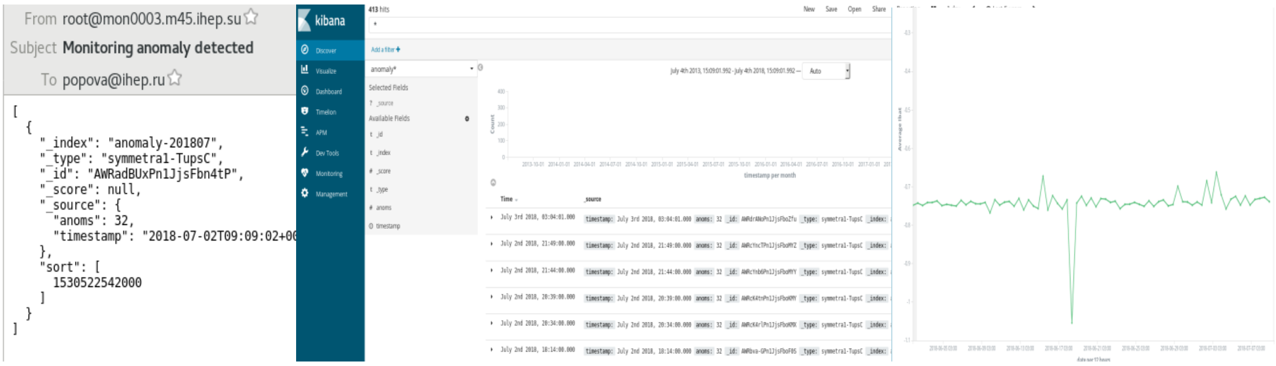


Figure 2. Collector for UPS system of the cluster [6]

Next, by cron, we extract metrics (Ibat, temperature) from ES and try to catch anomalous events using Twitter R DetectAnomaly package. If anomalies exist, we put them into ES and send email to administrators (Fifure 3).


From root@mon0003.m45.ihep.su

Subject Monitoring anomaly detected

To popova@ihep.ru

```
[
  {
    "_index": "anomaly-201807",
    "_type": "symmetra1-TupsC",
    "_id": "AWRadBUxPn1JjsFbn4tP",
    "_score": null,
    "_source": {
      "anoms": 32,
      "timestamp": "2018-07-02T09:09:02+00
    },
    "sort": [
      1530522542000
    ]
  }
]
```

Figure 3. Work results (mail, ES records in Kibana, anomalies on grapth)

5. Results

At this stage, we received a non-interactive script system for anomaly detection. It allows us to track the anomalous events in electricity supply system (Symmetra) in realtime. Well-timed identified anomalies prevent expensive equipment from damage and reduce downtime in the computer center. But we got many incorrect anomalies and false positives make our work inaccurate. Furthermore we analyzed only two params (Ibat and Temp) separately.

6. Conclusion

In this work we considered the notion of anomaly, anomaly types, data processing methods. Statistical methods are described more detailed, so they are most suitable for working with time series. Further, using the set of real-time scripts, the Ibat, Temp metrics are extracted from the ESK stack. These data are checked for anomalies using the Twitter R AnomalyDetection package. The found anomalies were returned to the ESK, the letter was sent to system administrators. The obtained results help in tracking problems in the equipment maintenance and operation, but they are numerous and inaccurate. In addition, the parameters are hardly defined. In future, we plan to test other statistical methods for data processing, consider indicators in terms of context and multidimensionality (for example, time as a context indicator, Ibat-Temp relationship), increase the number of metrics for analysis, make the analysis system interactive and more user-friendly for the end user.

References

[1] Varun Chandola, Vipin Kumar. Anomaly Detection: A Survey [DSG - Advanced Topics in Cyberphysical Systems]. Available at: https://www.vs.inf.ethz.ch/edu/HS2011/CPS/papers/chandola09_anomaly-detection-survey.pdf. (accessed 10.05.2018)

[2] Preetam Jinka, Baron Schwartz. Anomaly Detection for Monitoring, 2015, O'Reilly Media, Inc.

[3] Jordan Hochenbaum, Owen S. Vallis, Arun Kejariwal. Automatic Anomaly Detection in the Cloud Via Statistical Learning Twitter Inc. [arXiv.org e-Print archive]. Available at: https://arxiv.org/pdf/1704.07706.pdf. (accessed 10.05.2018)

[4] Introducing practical and robust anomaly detection in a time series [Introducing practical and robust anomaly detection in a time series]. Available at: https://blog.twitter.com/engineering/en_us/a/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series.html. (accessed 10.05.2018)

[5] What is Elasticsearch, and How Can I Use It? [What is Elasticsearch, and How Can I Use It?] Available at: https://qbox.io/blog/what-is-elasticsearch. (accessed 10.05.2018)

[6] V. Kotliar, V. Anshukov, V. Ezhova, V. Gusev, A. Kotliar, G. Latyshev, A. Shishov. Development of the active monitoring system for the computer center at IHEP // CEUR Workshop Proceedings. February 2017: Vol. 1787.- pp. 317-322