Towards Novel Statistical Methods for Anomaly Detection in Industrial Processes

Simone Tonini[1,*], Fernando Barsacchi[2], Francesca Chiaromonte[1,3], Daniele Licari[1] and Andrea Vandin[1,4]

[1]Department of Excellence EMbeDS, Sant'Anna School for Advanced Studies, Piazza Martiri della Libertà, 33, Pisa, 56127, Italy
[2]A.Celli Nonwovens S.p.A., Via Romana Ovest, 252, 55016, Porcari, Lucca, Italy
[3]Dept. of Statistics and Huck Institutes of the Life Sciences, The Pennsylvania State University, USA
[4]DTU Technical University of Denmark, Lyngby, Denmark

Abstract

These notes present a summary to [1], where we present an empirical data-driven methodology to identify anomalies (points anomaly) in industrial contexts where the production process is characterized by time series with unknown distribution. The methodology was developed within the AutoXAI2 project, born from the cooperation between Sant'Anna School for Advanced Studies of Pisa (EMbeDS – department of excellence for economics and management in the era of data science), and the A.Celli company of Lucca, leader in the supply of machinery and advanced technologies for the paper and nonwovens market (www.acelli.it). The project is co-funded by Tuscany region and the company. The objective of the project is to identify anomalies during the production process of the A.Celli tissue machine.

Keywords

anomaly detection, industrial processes, mahalanobis distance

1. Introduction

[1] introduces a method for anomalies detection in industrial contexts in the particular case where:

- the production process is characterized by time series with unknown distribution,
- the data have no labels of past anomalies (i.e., we are in an unsupervised framework).
- any information relating to the type of anomalies to be found is missing. In particular, there is no information about their duration and which variables should be monitored more frequently.

We propose an agnostic 5-steps methodology to classify one or more observations as anomalies in a context as that described above, which is based on first principles of statistical learning (variance inflation factor, Mahalanobis distance, and Chebyshev's inequality). The proposed methodology is easy to implement, fast to run, does not require the knowledge of the distribution of the variables, has low cost, and allows an easy interpretation of the results. However, it requires collaboration with a domain expert from the company, especially in the first phase, to assess that the identified anomalies are really relevant to the production process. In particular, short anomalies could be simple sensor noise, or there could be long events due to insignificant variables.

2. Methodology

Let $\mathbf{X}_A$ be an $n \times T_A$ array of observations relative to the production process without anomalies. We want to detect possible anomalies in $\mathbf{X}_B$, i.e. a new $n \times T_B$ array of observations, where $T_A \gg T_B$. For the $n \times T_{tot}$ matrix $\mathbf{X}_{tot} = (\mathbf{X}_A, \mathbf{X}_B)$, with $T_{tot} = T_A + T_B$, the proposed procedure is summarized as follows.

- *Step 1 - Data cleaning.* Remove the variables irrelevant according to the domain expert.
- *Step 2 - Smoothing.* For a window of size h, a median filter applied to each of the T_{tot}-dimensional vector $\mathbf{x}_i$'s works as follows:

$$\begin{aligned} w_{i1} &= med(x_1, \dots, x_h), \\ w_{i2} &= med(x_2, \dots, x_{h+1}), \\ &\vdots \\ w_{iT_{tot}-h+1} &= med(x_{T-h}, \dots, x_{T_{tot}}). \end{aligned}$$

  Thus, we replace the $n \times T_{tot}$ matrix $\mathbf{X}_{tot}$ with the $n \times T_{tot} - h + 1$ matrix $\mathbf{W}$.
- *Step 3 - VIF.* Get the $m \times T_{tot} - h + 1$ matrix $\widetilde{\mathbf{W}}$, where $m \leq n$ composed by variables with VIF<5.

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Simone Tonini
simone.tonini@santannapisa.it (S. Tonini); f.barsacchi@acelli.it (F. Barsacchi); francesca.chiaromonte@santannapisa.it (F. Chiaromonte); daniele.licari@santannapisa.it (D. Licari); andrea.vandin@santannapisa.it (A. Vandin)
https://orcid.org/0000-0002-6325-9533 (S. Tonini); https://orcid.org/0000-0002-0037-5457 (F. Barsacchi); https://orcid.org/0000-0001-5605-9886 (F. Chiaromonte); https://orcid.org/0000-0002-2963-9233 (D. Licari); https://orcid.org/0000-0002-2606-7241 (A. Vandin)

- *Step 4 - Anomalies detection.* Calculate $MD(\widetilde{\mathbf{W}})$, i.e. the Mahalanobis distance relative to $\widetilde{\mathbf{W}}$, and standardize it to obtain $Z_{md(\widetilde{\mathbf{w}})}$. Then, we obtain a new binary variable, named Y, equal to 1 if $|Z_{md_t(\widetilde{\mathbf{w}})}| \geq k, t = 1, \dots, T$, and 0 otherwise.
- *Step 5 - Variable detection.* Identify the variable that most contributes to the observed anomalies through

$$max_{1 \leq j \leq m} Corr(Y, X_{Bj}). \quad (1)$$

3. Validation

We apply our procedure to the Server Machine Dataset (SMD). SMD is a 5-week-long dataset collected from a large Internet company [2] and composed of 38 variables. It contains metrics like CPU load, network usage, memory usage, etc, and is made up of data from 28 different machines where the observations are collected per minute. In particular, the dataset contains periods with and without anomalies and therefore can be used to validate the proposed methodology. We try to identify anomalies for the first machine only and the main goal is getting a large $Precision = \frac{TP}{TP+FP}$, where TP and FP indicate true and false positives, respectively. The dataset for the first machine of the SMD dataset is composed as follows:

- X_A: 15800 timepoints without anomalies.
- $X_B, \dots, X_I$: 8 clusters of anomalies, about overall 12700 timepoints for all 8 clusters

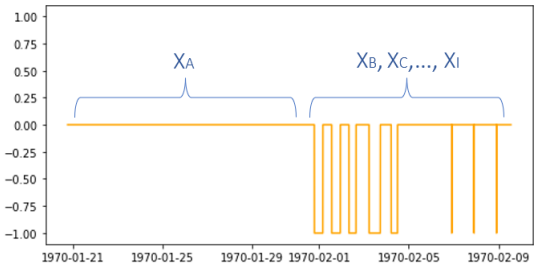

Figure 1: Anomalies in the test set of the first machine of the SMD dataset. The x-axis shows the dates on which the data were collected. Anomalies are denoted with y value, a zero-value denotes the absence of anomalies, while -1 indicates the presence of anomalies.

The proposed methodology is thus applied to the datasets (X_A, X_B), (X_A, X_C),..., (X_A, X_I), separately.

We consider three different values of the smoothing parameter h, namely 1 (no smoothing), 10 (medium smoothing), 60 (high smoothing). The results of this validation study can be summarized as follows.

- *No smoothing:* Without smoothing the data the proposed methodology always identifies the anomalies in correspondence with the true ones. However, there are also some *false positives* among the identified anomalies. It means that without a smoothing step, the proposed procedure is sensitive to noises in the data, which can be considered short irrelevant shocks. Therefore, with no smoothing, we get $Precision < 1$.
- *Medium smoothing:* By applying a slight/medium level of smoothing, it is observed that the problem of false positives is largely resolved. Furthermore, it is observed that smoothing makes our procedure more sensitive to true anomalies of long duration and consequently the number of true positives increases.
- *High smoothing:* In the most extreme considered case of high smooth, there are no false positives in any of the considered cases. However, we select to identify only anomalies with a relatively large duration. Therefore, with a large smoothing, we get $Precision = 1$.

References

[1] S. Tonini, F. Barsacchi, F. Chiaromonte, D. Licari, A. Vandin, Towards novel statistical methods for anomaly detection in industrial processes, in: Companion of the 2023 ACM/SPEC International Conference on Performance Engineering, ICPE '23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, p. 147–153. URL: https://doi.org/10.1145/3578245.3585036. doi:10.1145/3578245.3585036.

[2] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, D. Pei, Robust anomaly detection for multivariate time series through stochastic recurrent neural network, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & amp; Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2828–2837. URL: https://doi.org/10.1145/3292500.3330672. doi:10.1145/3292500.3330672.