Enhancing Datasets For Artificial Intelligence Through Model-Based Methods

Getting enough data to train useful AI models for industrial processes.

By Dirk Mayer and Ulf Wetzker

Industrial plants and processes are now digitized and networked, and AI can be used to evaluate the data generated by those facilities to increase productivity and quality.

Machine learning (ML) methods can be applied to:

- Product quality classification in complex production processes.
- Condition monitoring of technical systems, which is used, for example, in the diagnosis of drive systems, production plants, as well as for the wireless communication of critical automation components.
- Abnormality detection within sensor data or process information. Early initiation of countermeasures enables the reduction of system failures.
- Prediction of results or events based on preceding measurements.
- Optimization of production processes, taking into account the material flow.
- Training of intelligent robotic systems.

Some of the most popular machine learning applications are based on smartphone user data or sources from the Internet (social networks, Wikipedia, image databases, etc.) For the latter, very large training datasets are used, e.g., about 45 TB of text data for training the OpenAI GPT-3 [1].

Real industrial applications leverage much smaller data sets. This makes it difficult to train high-performance ML models and consequently to fully leverage the potential added value. The data sets are often incomplete for the following reasons:

- In industrial measurement campaigns, it's often not possible to comprehensively monitor all important conditions that need to be classified. This applies, in particular, to data from system faults or defects.
- It is frequently impossible to collect data on all units of a machine model, so mechanical or electronic differences and environmental influences (e.g., temperature fluctuations) are not reflected in the data.
- The data is commonly digitized, filtered, and compressed, so that information is lost.
- Consequently, data are also not completely labeled, i.e., assigned to a state to be classified later.

These incomplete data sets lead directly to over-fitted AI models and a lack of generalization. At the same time, measurement campaigns covering all possible variations are not economically feasible.

<u>Solutions</u>

To overcome these problems, insufficient data sets must be cleaned and augmented.

In industrial processes, data from time series play a particularly important role (e.g., sensor data, process parameters, log files, communication protocols). They are available in very different temporal resolutions – a temperature sensor might deliver values every minute, while for a spectral analysis of wireless network requires over 100 million samples per second.

The objective is to reflect all relevant states of the processes and uncertainties due to stochastic effects within the augmented time series. To add additional values to measured time series of an industrial process, insights into the process are beneficial. Such representation of the physical background can be called a model. In terms of model building, a division can be made into the following levels:

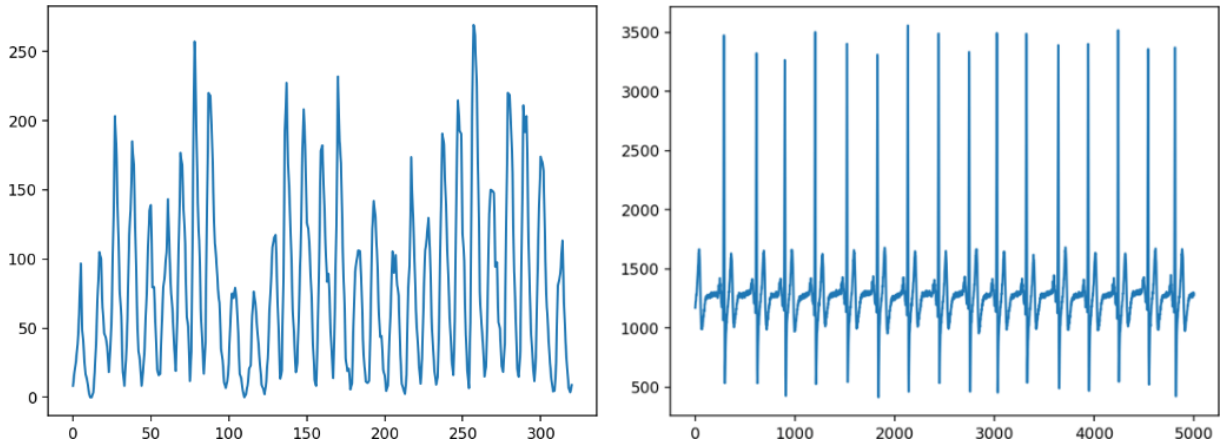
Black box model	Grey box model	White box model
Very little knowledge is available about the process and the signal (time series).	Signal shapes can be described and assigned to states by parameters. The parameters are not necessarily physically significant.	Complete parametric model of the physical process is available. Parameters are physically meaningful.

This allows us to derive strategies for a model-based generation of data. In order to generate longer and more appropriate time series for the training of AI models, the described strategies should be combined in an application-specific way:

- **Black Box.** Unsupervised learning can be used to generate artificial time series. This creates new, "similar" data sections without a deeper physical understanding of the waveforms. However, a relatively large amount of data is required and the relation between the sections is not physically motivated.
- **Grey Box.** Generation of sections in the time series from physical understanding, e.g., superposition with certain patterns belonging to relevant classes or distortion of measured time series. This requires numerous measurements and a basic understanding of which waveforms are assigned to which states or classes.
- **White Box.** Generation of time series from a system simulation, which theoretically does not require any measurements at all. In reality, however, completely white ("snow white") models are generally not possible, since parameters must always be matched with reality.

In the field of image processing, the augmentation of data might be intuitively easier. In contrast, the augmentation of time series mostly requires understanding of the underlying process. Depending on the depth of prior knowledge, model-based and synthetic data can be used. The optimal strategy for extending data usually follows economic considerations. Depending on the problem, collecting a complete set of measurement data, or generating physically meaningful models, can be very costly. In industrial practice, one will mainly use methods from the "grey box" category with limited experimental and analytical effort.

Interesting perspectives for interdisciplinary approaches also arise. Time series can be found in completely different processes even outside of technology and industry. The underlying processes are completely different, but the characteristics of the time series are very similar. In the picture below, two time series are shown, which have some similarity due to the oscillation of the values. However, they are generated by completely different processes. On the left is shown the periodic oscillation of the solar activity (period about 10 years, x-axis in years from the year 1700, sampling rate 1 year [2]). On the right the ECG of a human, period approx. 1s, sampling rate 1/300s [3]). This offers potential for the transfer of methods across domains, e.g., by using the sophisticated models for speech and text processing for the data augmentation in the medical domain [4].



In order to achieve a sustained increase in the performance of a trained model, it is necessary to incorporate human knowledge about the process. Methods from the field of human-in-the-loop ML, such as active learning, offer the option to move from a black-box approach to a grey-box model.

Currently, there is no systematic approach or simple tools for industrial applications that combine the above methods in a meaningful way to enable efficient data augmentation. This is the subject of current research.

References

[1] https://www.springboard.com/blog/ai-machine-learning/machine-learning-gpt-3-open-ai/

[2] https://wwwbis.sidc.be/silso/datafiles#total

[3] Richter. A: Entwicklung eines Systems zur Erfassung affektiver Zustände auf der Grundlage von Vitalparametersensordaten, Master Thesis, TU Chemnitz, July 2021.

[4] Bird, J. J., Pritchard, M., Fratini, A., Ekart, A., & Faria, D. R. (2021). Synthetic Biological Signals Machine-Generated by GPT-2 Improve the Classification of EEG and EMG Through Data Augmentation. IEEE Robotics and Automation Letters, 6(2), 3498-3504. https://doi.org/10.1109/LRA.2021.3056355