SMART PROCESS DATA ANALYTICS FOR MODEL PREDICTION

Weike Sun and Richard D. Braatz*
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract Overview

Data analytics are increasingly changing how manufacturers make critical decisions and designs. Although data analytics tools are valuable for improving the production process, the selection of the best method requires a substantial level of expertise. In practice, data analytics practitioners typically choose the methods based on familiarity or pick a method with the best cross-validation results from a large number of methods. Neither of these two approaches is optimal: either the method selection is strongly biased or the method can over-fit the data. This work presents a robust and automated smart process data analytics software which empowers the users to focus on goals rather than on methods, and automatically transforms manufacturing data into intelligence. The approach first applies tools to automatically interrogate the data to ascertain its characteristics (e.g., nonlinearity, multicollinearity, dynamics) to guide the methods selection for a particular application. This information is then used to select among the best-in-class data analytics method which is pre-selected based on our domain knowledge on process data analytics. The framework can be applied to model prediction, classification, and process monitoring. This poster focuses on prediction and the proposed approach is demonstrated in several case studies of manufacturing data.

Keywords

Manufacturing process, Process monitoring and diagnostics, Model prediction.

Introduction

Advances in machine learning and data science have enabled obtaining an in depth understanding of complex data. Advanced data analytics are transforming many fields, including manufacturing, especially when first-principles models are not available (Raghupathi and Raghupathi, 2014; Lee et al., 2014; Qin, 2014). Process data analytics (aka process analytics) encompass a wide variety of methods including chemometrics (Brereton, 2003), multivariate data analysis (Everitt and Dunn, 2001), time series analysis (Chatfield, 2016), and many others. Newer artificial intelligence (AI)-based tools such as machine learning (Monostori, 2003) are not yet extensively being applied in manufacturing processes, and deserve exploration and understanding so that regulators and industry alike can maximize their potential in risk-based decision making.

Process analytics can be applied for individual unit operations up to the level of the entire manufacturing process. They have the potential to be used by manufacturers to ensure product quality in a number of ways. They can be used to *make predictions*, such as providing modification procedures for downstream processes to improve product quality. They can be used for *control*, that is, to compute adjustments to the critical process parameters to move the critical quality attributes towards desirable values. They can be used to *improve*

* Corresponding author: Richard D. Braatz, braatz@mit.edu

process understanding, which can lead to improving troubleshooting capabilities during manufacturing. Finally, models can be used for the *real-time detection and diagnosis* of anomalous behavior, which is typically referred to as *anomaly detection* or *fault detection* in the literature (Chiang et al., 2000).

However, specific challenges are encountered in the application to manufacturing processes. The first challenge comes from the complexity of manufacturing processes. Oftentimes, the number of available samples is not adequate to fully describe the intrinsic varieties of the underlying processes. Moreover, for some applications, only limited samples are available. Another challenge is that the signal-to-noise ratio of the data can be low and some latent factors are not measureable. Furthermore, a substantial level of expertise is required to select the best process data analytics method for a specific application, to avoid overfitting and to ensure that an accurate process understanding is obtained. But in reality, it is hard for the data analytics practitioners to be knowledgeable about all techniques. In actual practice, users typically default to using familiar techniques or simply trying all models. Neither of these approaches is optimal.

This work proposed a robust and automated approach for data analytics model selection and fitting for the problem of model prediction, which allows the user to focus on specifying the objectives of the process data analytics rather than spending extensive time and effort in learning and selecting among methods. The basic concept was pioneered by Severson et al. (2018), which takes a bottom-up approach that makes the decision about which technique to use based on the characteristics of the data and available domain knowledge. To be best of our knowledge, this software is the first for automating data analytical model selection and fitting for manufacturing processes. The selection procedure and the candidate models incorporate artificial intelligence and expertise in process data analytics and domain knowledge on manufacturing, and are designed to provide models of high robustness, accuracy, and interpretability.

Methodology

The basic approach first applies tools to automatically interrogate the data to ascertain their key characteristics for model prediction, including nonlinearity, multicollinearity, and dynamics. This information is then used to select a best-in-class process data analytics tool, which can be graphically illustrated in the form of a triangle (Figure 1). The full implementation involves a more detailed decision tree for final model selection of a particular dataset with high-quality cross-validation, designed to achieve the objectives specified by the user.

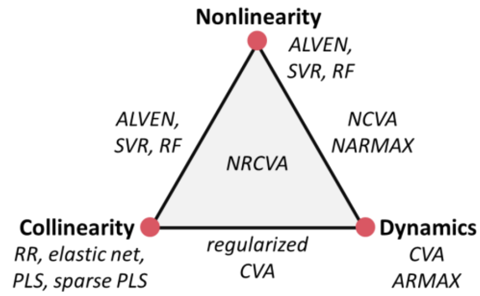

Figure 1. Data analytics triangle for model prediction, which maps modeling techniques to data characteristics: ALVEN = algebraic learning via elastic net, ARMAX = autoregressive moving average with exogenous input, CVA = canonical variate analysis, PLS = partial least squares, RF = random forest, RR = ridge regression, SVR = support vector regression, NARMAX = nonlinear ARMAX, NCVA = nonlinear CVA.

Conclusion

A robust and automated approach for model selection and fitting is proposed for model prediction in manufacturing processes, which is based on the specific application scenario and data availability. The smart process analytics approach considers important data characteristics and uses this information for method selection. The user's objectives are also taken into consideration during method selection. The approach facilitates consistent application of best practices and continuous improvement of tools and decision making. Case studies involving real manufacturing datasets demonstrate the effectiveness of the proposed method, which outperforms the methods commonly applied in industries.

Acknowledgments

This work is supported by the U.S. Food and Drug Administration. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the financial sponsor.

References

Brereton, R. G. (2003). *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. John Wiley & Sons, Chichester, United Kingdom.

Chatfield, C. (2016). *The Analysis of Time Series: An Introduction.* Fifth edition, Chapman and Hall, London.

Chiang, L. H., Russell, E. L., & Braatz, R. D. (2000). *Fault Detection and Diagnosis in Industrial Systems*. Springer Verlag, London.

Everitt, B. S., & Dunn, G. (2010). *Applied Multivariate Data Analysis*. Second edition, Wiley, New York.

Lee, J., Kao, H. A., & Yang, S. (2014). Service innovation and smart analytics for Industry 4.0 and big data environment. *Procedia CIRP*, 16, 3-8.

Monostori, L. (2003). AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing. *Engineering Applications of Artificial Intelligence*, 16(4), 277-291.

Qin, S. J. (2014). Process data analytics in the era of big data. *AIChE Journal*, 60, 3092-3100.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2, 3.

Severson, K. A., VanAntwerp, J. G., Natarajan, V., Antoniou, C., Thömmes, J., & Braatz, R. D. (2018). A systematic approach to process data analytics in pharmaceutical manufacturing: The data analytics triangle and its application to the manufacturing of a monoclonal antibody. In *Multivariate Analysis in the Pharmaceutical Industry*, edited by A. P. Ferreira, J. C. Menezes, and M. Tobyn, Elsevier, Chapter 12, 295-312.