SMART PROCESS DATA ANALYTICS FOR SUPERVISED CLASSIFICATION

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

Abstract Overview

Many powerful tools for data analytics and machine learning have become developed, with especially rapid growth in open-access and commercial software in the last decade. This situation has motivated the development of *smart data analytics*, that is, a decision tree that automatically selects the best methods based on a systematic interrogation of the characteristics of the dataset. The detailed development of the decision tree depends on the objective of the data analytics application. This work designs smart process data analytics for the objective of classification among datasets with predefined labels. This supervising learning problem arises for manufacturing processes in several contexts, including fault diagnosis. Most of the decision tree can be visualized in form of a data analytics triangle, in which each part of the triangle is associated with one to three characteristics: normality, nonlinearity, and dynamics. The potential improved classification performance obtained by using this fully automated approach to data analytics method selection is demonstrated in case studies.

Keywords

Data analytics, Supervised learning, Supervised classification, Machine learning, Fault diagnosis.

Extended Abstract

Many powerful tools for data analytics and machine learning have become developed, with especially rapid growth in open-access and commercial software in the last decade. Becoming an expert in the selection and application all of these methods is challenging, as methods come from many disciplines, including applied statistics, analytical chemistry, pattern recognition, operations research, and computer science (e.g., Chiang et al., 2001). No single method produces the best results for all real-world datasets, and in practice either a sub-optimal method or subset of methods favored by the data analyst is applied or a significant amount of time is spent comparing results from many different methods to try to find the method that produces the best results. This situation has motivated the development of *smart data analytics*, that is, a decision tree that automatically selects the best methods based on a systematic interrogation of the characteristics of the dataset (Severson et al., 2018).

The detailed development of a decision tree depends on the objective of the data analytics application. Severson et al. (2018) developed a decision tree for the objective of regression, that is, the construction of a mathematical model that predicts process outputs based on process inputs. In contrast, this work considers the alternative objective of supervised classification. More specifically, smart process data analytics is designed for the objective of classification among datasets with predefined labels. This supervising learning problem arises for manufacturing processes in several contexts.

An example of such a problem is the classification of datasets of operating variables collected in the first batch unit operation in a series of batch unit operations based on

* To whom all correspondence should be addressed

whether the final product is in-spec or out-of-spec. A model can be developed based on a training data set to construct a binary classifier predicting whether the final product will be in-spec or out-of-spec. When new data on operations are collected, the supervised classifier constructed from past data can be used to assess whether the final product is likely to meet specifications. If there is a very high likelihood of the material eventually resulting in poor quality product, that information can be used to decide to dump material from the first batch, rather than moving the material through a series of expensive downstream unit operations to eventually learn that the product at the end is out-of-spec.

Another example application of supervised learning for a manufacturing process is to assess archival data into classes that represent normal operating conditions and different faulty conditions observed in the past. Assuming that these archival data can be labeled successfully, it is possible to create a supervised classifier. When new data on operations are collected, the supervised classifier is used to assess whether the operations are likely to be normal or be associated with past observed faults. According to these results, faults can be detected and measures can be taken to overcome the faults.

As in the regression problem as observed by Severson et al. (2018), most of the decision tree can be visualized in form of a data analytics triangle. Each part of the triangle is associated with a different data analytics method, so that the data integration and the data analytics triangle collectively guide the user to the best data analytics method to apply for a particular dataset. For supervising classification, the three corners of this triangle represent the algorithms suitable for each of three characteristics: normality, nonlinearity, and dynamics. The edges of the triangle show the supervised learning algorithms that can be used if two of the three characteristics are important for the problem at hand, while the center of the triangle combines all three characteristics. Interrogation of the data in the proposed manner enables the user to identify a supervised learning algorithm that is best for the given classification problem.

Additional to the visualization of the decision tree, an interrogation framework is proposed for determining the extent to which each of the factors normality, nonlinearity, and dynamics are important for the given data set. Based on the framework by Severson et al. (2018), different techniques are proposed and adapted in order to measure the relevant characteristics for the classification case. From these results, the best algorithm can be selected from the proposed data analytics triangle.

Generally, it is well understood that an algorithm that does not consider a certain characteristic present in a given dataset is going to perform poorly because it is unable to capture the relationship between the model inputs and outputs. Some researchers have proposed to always apply some method such as dynamic neural networks that can handle all of the potential characteristics in the dataset. The drawback of such general methods is that datasets rarely contain all of the characteristics (e.g., non-normality, nonlinearity, dynamics) that can occur in process datasets, and that generality results in the method being prone to overfitting when it tries to construct types of relationships that are not existent in the given dataset. Consequently, better performance is obtained by data analytics methods that only consider the relevant characteristics in the data.

The smart data analytics for supervised classification is illustrated in case studies, which compare the misclassification rates of the selected supervising learning methods with alternatives. Significant improvement in classification performance is demonstrated in a fully automated procedure for data analytics method selection.

Acknowledgments

Financial support from the U.S. Food and Drug Administration is acknowledged.

References

Chiang, L. H., Russell, E. L., Braatz, R. D. (2001). Fault Detection and Diagnosis in Industrial Systems. *Springer-Verlag*, London, UK.

Severson, K., 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.