PSIG 1511

Advanced Analytics for Continuous Emission Monitoring Systems

Vadim Shapiro, Statistics & Control, Inc., Dmitriy Khots, iMath Consulting

This paper was prepared for presentation at the PSIG Annual Meeting held in New Orleans, Louisiana, 12 May – 15 May 2015.

ABSTRACT

Government regulatory bodies, such as the United States Environmental Protection Agency (EPA), set forth regulations for emissions. The Code of Federal Regulations Title 40 Protection of Environment part 60—Standards of Performance for New Stationary Sources—and part 75—Continuous Emission Monitoring—are the two EPA federal regulations that guide the design of all Continuous Emission Monitoring Systems (CEMS). CEMS, especially extractive CEMS, typically have three major components: sample transport and conditioning, sample analysis, and data acquisition and storage. This paper focuses on advanced analytics that can be used to provide reliable and accurate service, even when extractive CEMS is off-line, data are missing, or if a monitoring system is not installed.

The advanced analytics consists of three components: emission measurement data produced and stored by the extractive CEMS or manually entered laboratory test data, a model that relates emission measurements to the process operating mode, and a calculated optimal operating mode that minimizes the emissions while keeping the client's operation economically feasible. This paper describes the data used for the advanced analytics, the two modeling techniques used to predict continuous proportions of exhaust gas components of interest, and the methodology for optimizing the interdependence of the exhaust gas component proportion with other process variables.

INTRODUCTION

Advanced analytics represents a suite of tools designed to work in conjunction with Continuous Emission Monitoring Systems (CEMS) and in particular with extractive CEMS to ensure regulatory emission requirements outlined by the United States Environmental Protection Agency's (EPA) Code of Federal Regulations (CFR) Title 40 Protection of Environment, parts 60 and 75. These advanced analytics integrate with any legacy extractive CEMS as long as the legacy system is fully compliant across its three major components: sample transport and conditioning, sample analysis, and data acquisition and storage.

Advanced analytics identify mathematical dependencies between various process variables and the turbine/boiler exhaust gas composition produced by the extractive CEMS or by laboratory tests. These dependencies create models, which predict how the exhaust gas composition will respond to changes in the operating mode. Ultimately, the optimal operating mode is found based on the minimum exhaust criteria.

Advantages of analytics include ensuring regulatory compliance, environmental safety, and plant economic feasibility, as well as providing an additional layer of redundancy to existing extractive CEMS, 99.9% uptime, and integration with the plant's HMI for comprehensive alarming and reporting.

REGULATORY CONSIDERATIONS

The EPA is the federal body that governs the requirements for all CEMS. The CFR 40 (Protection of Environment) part 60 (Standards of Performance for New Stationary Sources) and part 75 (Continuous Emission Monitoring) are the two federal regulations that guide the design of all CEMS implementations. Part 60 predominantly deals with emission guidelines and performance standards across various industries and production processes. Part 75 specifically deals with continuous emission monitoring, which is the foundation for the U.S. "cap and trade" program.

These regulations apply to the processes that occur in the extractive CEMS, including continuous sampling of the exhaust gas from a processing point; sample filtering and conditioning (typically dependent on the nature of the gas); selecting gas analysis techniques, such as spectroscopic absorption, luminescence, electroanalysis, electrochemical analysis, and paramagnetism; and data storage and quality management (e.g., handling of missing data).

Advanced analytics align with these requirements by providing a reliable and accurate service, even when the extractive CEMS is off-line or unavailable (because emission values can now be predicted based on other process variables, i.e., these predicted values can be used in place of missing data). In particular, if a monitoring system is not installed and real-time data are not available for model construction, the advanced analytics can use manually entered lab test analysis data. Data can be obtained once a week, once a month, or even less frequently, and the model will still be calibrated.

Additionally, advanced analytics provide redundant analyses to ensure accuracy as well as a substantial model testing. The algorithmic approach to accuracy and redundancy testing is described in detail below.

ANALYTICS ARCHITECTURE

The advanced analytics architecture consists of three components: emission measurement data produced and stored by the extractive CEMS (or manually entered laboratory test data), the model that relates emission measurements to the process operating mode, and the optimal operating mode that minimizes the emissions while keeping the client's operation economically feasible.

Data

Data come from three major sources: extractive CEMS data (preferred), laboratory testing data (whenever extractive CEMS is off-line or does not exist), and process data. Signal processing techniques filter and transform input data to prepare it for modeling.

In situations where the model has to use human-entered data (e.g., laboratory test data), data are often not clean or are missing. The advanced analytics deal with such situations with the following tools: time series extrapolation, statistical imputation (using the mean or the median to replace missing values), and tree imputation methods (where auxiliary models are built to find relationships between missing data and other process variables). In extreme cases where the proportion of observations with missing values is too high, the variable is converted into a binary indicator according to equation (1), which presents additional predictive information to the model.

$$x_t = \begin{cases} 1 | x = missing \\ 0 | x \neq missing \end{cases}, \quad (1)$$

where x is the original variable. All data are automatically tested for unusual observations, or outliers, using the six sigma approach. Given m observations of an independent variable, $x(t_1),\ldots,x(t_m)$, the standard deviation σ is calculated using equation (2).

$$\sigma = \sqrt{\frac{\sum_{j=1}^{m}\left(x(t_j)-\bar{x}\right)^2}{m}}, \quad (2)$$

where $\bar{x} = \frac{\sum_{j=1}^{m} x(t_j)}{m}$ is the sample mean. Then an outlier, xO, is defined according to equation (3).

$$\begin{aligned} & x(t_o) \equiv xO \ iff \ x(t_o) > \bar{x} + \delta \cdot \sigma \\ & or \ x(t_o) < \bar{x} - \delta \cdot \sigma, \end{aligned} \quad (3)$$

where $\delta > 0$ is a sufficiently large parameter.

If any outliers are found, the data are transformed using max normal transformations, such as the natural logarithm or power transformations, thereby removing the biases introduced by these outliers.

Model

The advanced analytics primarily use two modeling techniques to predict continuous proportions of exhaust gas components of interest. These techniques are polynomial regression using ordinary least squares (OLS) fitting methodology and artificial neural networks utilizing multilayer perceptron (MLP) structure. The model is trained in online mode using a continuous champion/challenger approach, where the current (champion) model is checked at every time scan against a voting system of challenger models to ensure sustained model lift and to prevent model decay.

Given the target variable y (exhaust gas component proportion) and a set of n independent potential predictor variables $x_1,\ldots,x_n$ (process variables) with m observations measured in a time interval T, the polynomial regression algorithm estimates coefficient vector $\boldsymbol{\beta}$ that minimizes equation (4).

$$S = \sum_{j=1}^{m}\left(y(t_j) - P\left(\beta, x_1,(t_j),\ldots,x_n(t_j)\right)\right)^2 \quad (4)$$

This minimization is achieved by calculating partial derivatives of S with respect to each element $\beta_i \in \beta$, with $i = 1,\ldots,M$ given by equation (5).

$$\frac{\partial S}{\partial \beta_i} = 2\sum_{j=1}^{m}\left(y(t_j) - P\left(\beta, x_1,(t_j)\right)\right)\cdot \frac{\partial\left(y(t_j) - P\left(\beta, x_1,(t_j),\ldots,x_n(t_j)\right)\right)}{\partial \beta_i}, \quad (5)$$

Then, partial derivatives are set to zero, resulting in a system of linear equations shown in equation (6) that is solved using Cramer's rule.

$$\begin{cases} \frac{\partial S}{\partial \beta_1} = 0 \\ \vdots \\ \frac{\partial S}{\partial \beta_M} = 0 \end{cases} \quad (6)$$

The resulting solution (if it exists) is the vector $\boldsymbol{\beta}$ that minimizes the distance from the fitted values of y to actual values of y and can be used for processing by the optimization submodule.

For artificial neural networks, the algorithm utilizes an MLP structure with the back propagation algorithm. The initial model input is a set of potential normalized predictor variables, $x_1,\ldots,x_n$, along with their weights, $w_{1j},\ldots,w_{nj}$, chosen at random for each neuron. Each neuron processes the inputs to produce an output that is the weighted sum of the inputs provided by equation (7).

$$h_j = \sum_i w_{ij} x_i \quad (7)$$

Each neuron's output is then activated using an activation function F, which could be a hyperbolic tangent or a logistic sigmoid function. The output is then shown by equation (8).

$$o_j = F\left(\sum_i w_{ij} x_i\right) \quad (8)$$

Next, each neuron's output becomes an input for the next neuron in the network, with corresponding weights assigned at random for the first run. Ultimately, the network's total output is the activated linear sum of output from all neurons from the last hidden layer, which can be expressed by equation (9).

$$\hat{y} = F\left(\sum_{i,j} v_{ij}\, g(x_i)\right), \quad (9)$$

where j runs across all units in the network, v_{ij} are the weights associated with the output from neurons in the last hidden layer, and functions g are a composite of all prior activation functions.

Figure 1 displays this process for an individual neuron with three inputs.

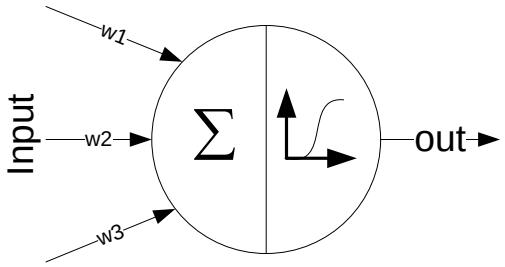

Figure 1 – Artificial neuron with three inputs

The next step of the algorithm evaluates the cost function shown in equation (10) across all observations in the training dataset.

$$C = \frac{1}{M}\sum_{p=1}^{M}\left(y(t_p) - \hat{y}(t_p)\right)^2 \quad (10)$$

The algorithm then adjusts all of the weights in the network so that C is minimized, which is accomplished with the gradient descent algorithm. The algorithm output is the set of weights, $\boldsymbol{W}$, across the entire network that are used by the optimization submodule.

The advanced analytics varies model parameters to ensure the most accurate output. The best model is selected using the generalized R^2 goodness-of-fit criterion tested on multiple cross-validation samples. Specifics of this selection are described in System Testing.

Optimization

For operating mode optimization problems, advanced analytics allow the plant operator to select decision variables (e.g., unit start and stop times, product flow, etc.), constraints (e.g., fuel gas, operating times, product flow, exhaust gas composition proportions, etc.), and the objective function (e.g., fuel gas, exhaust gas composition proportions, etc.). Note that depending on the optimization problem, a single attribute could be used either as a decision variable, a constraint, or an objective function.

The most common optimization problems involving emissions involve finding the optimal operating mode that either minimizes exhaust gas component proportions (e.g., SO_2, CO_2, etc.) subject to plant output constraints or maximizes plant output subject to specific exhaust gas component values. Thus, the exhaust gas component proportion is the objective function in the former case and a constraint in the latter case. In either case, it is crucial to have a function that accurately describes the interdependence of the exhaust gas component proportion with other process variables. Advanced analytics solve this problem using the modeling techniques already described.

As an example, if y is set to be the proportion of component of interest and $\hat{y}$ is the modeled value, then the first-case optimization problem can be stated according to equations (11) through (13).

$$\boldsymbol{min}\ \hat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{x_1},\ldots,\boldsymbol{x_n}) \quad (11)$$

subject to

$$c\_min_{x_i} \le x_i \le c\_max_{x_i}, \quad (12)$$

$$|O(x_1,\ldots x_n) - Out| \le \varepsilon, \quad (13)$$

where c_min_{xi} is the lower bound for every x_i, c_max_{xi} is the upper bound for every x_i, $O(\cdot)$ is the function that relates plant output to process variables, *Out* is the required output value, and $\varepsilon > 0$ is a sufficiently small parameter. The second-case optimization problem is then stated according to equations (14) through (16).

$$\min \hat{O} = O(x_1, \ldots, x_n) \tag{14}$$

subject to

$$c\_min_{x_i} \leq x_i \leq c\_max_{x_i}, \tag{15}$$

$$|f(x_1, \ldots, x_n) - Exhaust| \leq \delta, \tag{16}$$

where *Exhaust* is the required exhaust gas component proportion value and $\delta > 0$ is a sufficiently small parameter.

The advanced analytics choose the appropriate optimization solution method depending on the nature of the objective and constraint functions as well as decision variables. For example, in load allocation problems where decision variables are binary (unit shutdown/start-up), genetic algorithms will be used to find the optimal combination. The optimal operating mode is calculated at every time scan to accommodate the environment, process, and business changes, thereby ensuring the optimal solution for current conditions.

ALARMING

The advanced analytics provides the following alarms: data failure and emission outside of normal. Data failure alarms typically correspond to communication outages (i.e., when a specific measuring device stops sending signal values) or to incorrect signals (i.e., when there is a systematic error with the measuring device). Gross error detection algorithms are applied to determine one of four general data failure situations: bias, drifting, precision degradation, or complete failure. Thus, the analytics will generate four types of alerts based on the data failure type. Additionally, any corrective action taken will be communicated along with the alarm.

The advanced analytics applies six sigma process control techniques to generate alarms whenever abnormal emission values are detected. Figure 2 shows that an alarm will be generated whenever the proportion of the exhaust gas component of interest is in the zones indicated by LSL and USL (represented by the solid red lines).

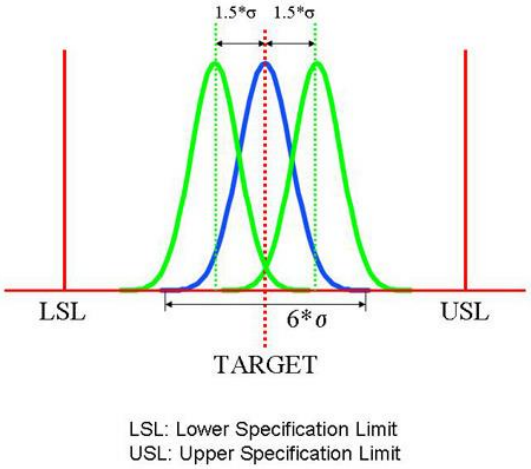

Figure 2 – Abnormal value detection

One of the major benefits of the analytics models is the fact that not only are the measured values being tracked for normal behavior but also the predicted exhaust gas component proportions are tracked. Therefore, it can be predicted whether emissions are outside the norm and, more importantly, when it will happen. Thus, plant operators can take proactive corrective actions to ensure compliance before any undesirable outcomes occur.

SYSTEM TESTING

The advanced analytics include self-tuning and redundancy algorithms that ensure the most accurate emission predictions. A brief synopsis of these algorithms is provided.

Let *y* be the exhaust gas component being predicted using the advanced analytics, $x_1, \ldots, x_n$ be the set of *n* uncorrelated and possibly transformed process variables that were deemed as predictive by the analytics model, and let $R^2_{champion}$ be the goodness-of-fit statistic associated with the current model in production (champion model), *f*. In the most general form, $R^2_{champion}$ is defined by equation (17).

$$R^2 = 1 - \left(\frac{L(0)}{L(\hat{\theta})}\right)^{2/m}, \tag{17}$$

where $L(0)$ is the likelihood of an intercept-only model (i.e., where no x_i's are significant), $L(\hat{\theta})$ is the likelihood that at least one x_i is a significant predictor, and *m* is the sample size. For example, in the case of ordinary least squares regression, $R^2_{champion}$ is given by equation (18).

$$R^2_{champion} = R^2_{adj} = 1 - \frac{\sum_j (y_j - f_j)^2}{\sum_j (y_j - \bar{y})^2} \frac{m-1}{m-n-1} \tag{18}$$

At every time increment, the advanced analytics generate a set of challenger models by varying the following parameters:

- The number of attributes used in the model, n
- The sampling time frame, T
- The number of cross-validation samples (bootstrapping), B
- The degree of the regression, p
- The number of interaction terms in the regression, Q
- The number of hidden layers in the artificial neural network, l
- The activation function in the artificial neural network, L
- The number of neurons in each layer, lN

Each challenger model goodness-of-fit is evaluated against the champion on cross-validation samples. The champion model remains whenever $\left|R^2_{challenger} - R^2_{champion}\right| \leq \varepsilon$, where $\varepsilon > 0$ is a sufficiently small parameter. The challenger model succeeds the champion whenever $\left(R^2_{challenger} - R^2_{champion}\right) > \varepsilon$. This technique ensures accuracy of the champion model as well as provides redundant solutions.

The existing CEMS and advanced analytics should also be subject to third-party RATA testing at least on a semiannual basis.

BIOGRAPHIES

Dr. Dmitriy Khots is an advanced analytics, mathematics, and optimization professional with over twelve years of experience across a broad spectrum of industries, including oil and gas, power generation, communications, finance, healthcare, and government. He currently has a dual role as the Vice President of Strategic Analytic Insight at West Corporation and as the President of a boutique consulting firm iMath Consulting, where Dr. Khots provides thought leadership to companies in the advanced process control space. Dr. Khots has a Ph.D. in mathematics from the University of Iowa, holds three patents, and has 40+ publications in theoretic and applied fields of mathematics as well as data mining and statistics.

Vadim Shapiro is co-founder and President of Statistics & Control, Inc. He manages all aspects of engineering, including developing the company's *OptiRamp*® Advanced Process Control software. Shapiro has over 20 years of leadership experience in systems and software engineering, engineering project management, and software product development in the fields of turbomachinery control, advanced process control, and power management systems. He holds five patents in the areas of turbomachinery and advanced process control, with several applications pending.