Online Learning for Run-by-run Prediction - James F. Bramante

[Enter co-author's name/co-author's name here]
James.Bramante@inficon.com
INFICON
70 Bridge St, Suite 202, Newton, MA 02458 USA
Phone: +1 -617-527-1219 Fax: +1-617-527-0917

Motivation

Real time and run-by-run prediction of faults and adverse maintenance events is a difficult task. Due to the speed at which wafers are processed, there is limited time between wafers during which dozens of status variable traces must be processed and used to predict future performance and perform interdiction, if necessary. Thus, prediction models must be trained online, one data point at a time, instead of fitting a model to the entire time series post hoc. Additionally, time series of maintenance indicators are often non-stationary (Fig. 1), such that their behavior, periodicity, and trend change over time. In fact, engineers often require predictions to be their most accurate when large changes in behavior occur, as these indicate degrading equipment performance. Thus, any predictive model must be able to adapt quickly to changes in the underlying generative process. Third, engineers are often interested in multiple prediction horizons. Unfortunately, no single prediction model is likely to fulfill all of these requirements. Instead, new methods are needed that combine models.

For the past few years, the state of the art forecasting models have achieved their results using ensembles of separately trained forecasting algorithms. International forecasting competitions have demonstrated that ensembles of relatively simple statistical and machine learning methods outperform most deep learning models in univariate time series prediction tasks for a wide range of forecasting horizons (Makridakis et al., 2018). A recently developed deep learning architecture, N-BEATS, is challenging the primacy of these methods, but still relies on ensembles of separate models (Oreshkin et al., 2020).

Unfortunately, these state-of-the-art techniques do not transfer to tasks like run-by-run fault prediction. The above techniques assume that time series are generated by a stationary process that can be characterized over multiple passes of the data. Online learning precludes the use of these methods, and maintenance indicators have non-stationary generative processes that cause these methods to fail as soon as the generative process changes.

Approach

We present here a method of online univariate forecasting using an ensemble of models whose weights are adjusted with the forecast horizon to improve the prediction accuracy of maintenance indicators and failure likelihood over multiple time horizons.

Our method builds on recent work to train SARIMA models online without specifying the orders of the model components a priori. In particular, our innovations are to include non-ARIMA models and an ensemble averaging approach for multiple time horizons. Our meta-algorithm cultivates a collection of models, or “experts”, for which it updates sets of weights indicating relative trust for each model at each of several time horizons. The trust weights are updated according to individual model loss during an n-ahead prediction task, where n is the time horizon for that set of weights. During prediction, the optimal weights for a given time horizon are predicted from the weights of existing time horizons. The trust weights are then converted to ensemble weights to produce the ensemble prediction from the full set of models.

Results

Trajectories of ensemble weights in our novel ensemble method demonstrates that our models clearly adjust for both time horizon and nonstationarity in the time series. Changes in periodicity, trend, and linearity of the data are accompanied by shifts in the model ensemble to emphasize one or a few models over the others (Fig. 2). Our meta-algorithm produces more consistently accurate predictions than any one of its component models over the entire time series.

Evaluated over 10 synthetic maintenance predictor time series with non-stationary behavior and 5 different sets of “experts”, we demonstrate that our model improves on the previous online ensemble method for varying time horizon (Fig. 3).

We demonstrate that an online ensemble modeling system can be used for accurate univariate prediction tasks without a priori knowledge about time series properties such as seasonality, variability, trend, or auto-correlation. These properties make this system ideal for predicting time to fault or time to failure on process data-based maintenance indicators.

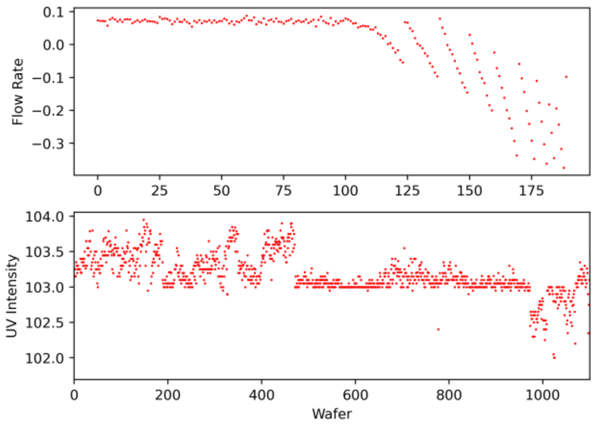

Figure 1. Two examples of maintenance indicators with non-stationary periodicity, trends, and abrupt signal degradation.

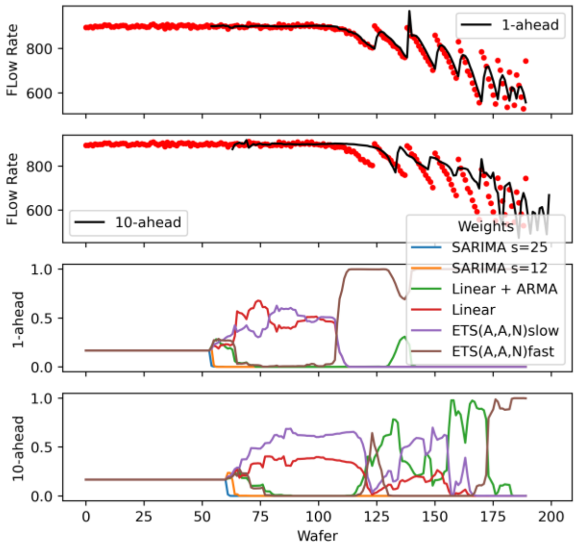

Figure 2. An example of the ensemble modeling system applied to predict the Flow Rate maintenance indicator in Fig. 1. The top two panels show 1-ahead and 10-ahead predictions, while the bottom two panels show the trajectories of ensemble model weights over the same time.

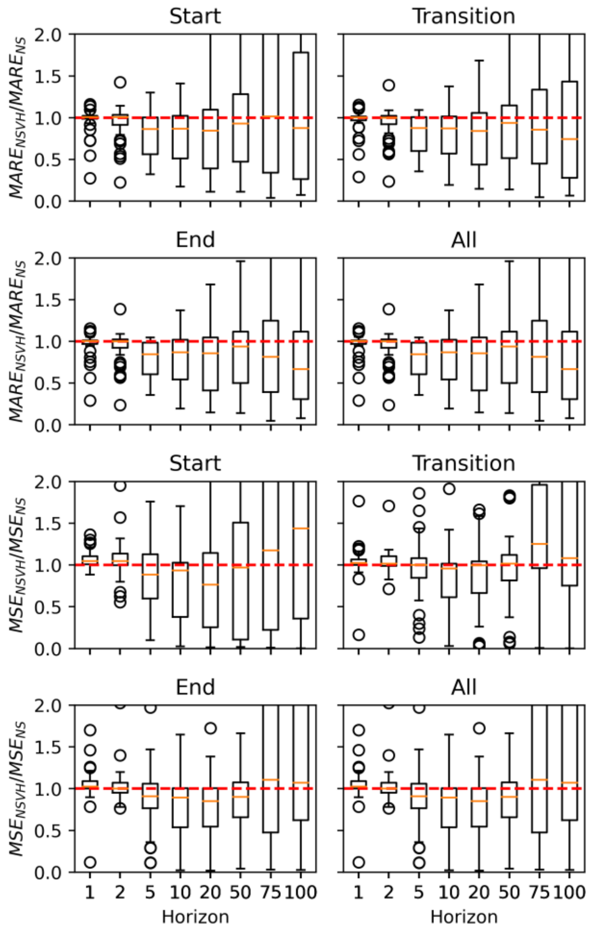

Figure 3. Comparison of the previous state-of-the-art online learning-based ensemble model (NS) with the variable-horizon version presented here (NSVH). The distributions of the ratio of NS to NSVH for two different loss functions, Mean Absolute Relative Error (MARE) and Mean Squared Error (MSE). Values below one (red dashed line) indicate NSVH outperforms NS. The loss functions were calculated separately for four different regions of synthetic maintenance indicators: Start – stationary behavior; Transition – 100 wafers after a sharp change in generative process; End – 100 wafers beyond Transition; All – for the entire time series. Each distribution represents 5 sets of experts applied to 10 different time series, demonstrating that NSVH performance is greater than NS across time horizon, ensemble composition, and region of the time series.