

To retrain or not to retrain – data cleaning vs. online learning in energy forecasting

Gregory Merkel and Richard Povinelli

In academic forecasting it is often an unstated assumption that forecasting models are retrained for every individual forecast to account for recent trends. However, in many domains, the immediate on-line data quality may not be good enough to support new models for each forecast point, and models are instead retrained monthly or annually. This paper seeks to determine which of these retraining schedules works best for short-term load-forecasting of natural gas.

Natural gas is the fuel of choice for many applications, including space heating and industrial processes. It is becoming the number one fuel for electricity generation. One of the problems that must be overcome in building accurate natural gas forecasters is the issue of data quality. Currently, when we build models a substantial effort is made to clean the data by removing outliers, identifying naive disaggregation, and correcting conversion factors. For models trained once a year this process is justified, and adequate time is available for data cleaning. However, for models retrained daily, this level of data cleaning is not possible.

We compare two forecasting strategies. The first involves models retrained yearly with extensive off-line cleaning and some data correction as needed. The second involves models retrained daily with rudimentary cleaning of on-line data and some seasonality adjustments. The tradeoff is that the models retrained yearly have greater opportunity for data cleaning and correction but can't incorporate recent trends, while models retrained daily can incorporate recent trends but some of the 'recent trends' may be due to data quality issues. For both strategies, we consider linear regression, artificial neural network, and ensemble models. The primary metric considered is weighted mean absolute percent error (WMAPE). We compare the forecasting strategies on more than 60 data sets.