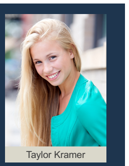
Use of deep neural networks to predict water quality variables at a decentralized wastewater treatment facility (E2.9)

Energy and resource recovery

Re-Inventing the Nation's Urban Water Infrastructure (ReNUWIt)


Taylor Kramer

Taylor Kramer, Kathryn Newhart, Tzahi Cath

Colorado School of Mines

Background

- Water and wastewater treatment facilities across the US will have to comply with lower effluent water quality limits. Meeting the new standards will require closer monitoring of effluent water quality; however, water and wastewater data are nonlinear, dynamic, and time-dependent; and a traditional, first-principals approach is not sufficiently accurate or timely for the purpose of process monitoring.

- Scientists have proposed the applications of data-driven machine learning and artificial intelligence (AI) to predict water quality for over a decade [1]. One of the largest unanswered questions concerning the world of AI remains is *'how much data?'* There are problems with insufficient data for machine learning models to learn how to predict outcomes accurately. Conversely, too much data is also a problem, leading to overfitted models (Figure 1).

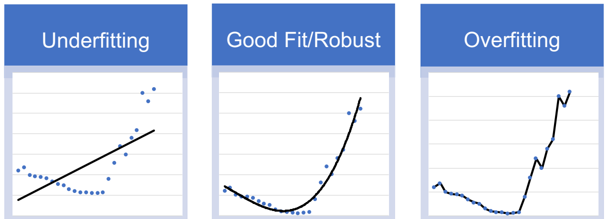


Figure 1. A visual representation of how a predictive model can be under- or over-fit if too little or too much data is used to train a NN model

- Multivariate data can be modeled by a deep neural network (DNN) estimator. Estimators represent a complete model and they can be created, trained, and used to evaluate the accuracy of a machine-learning model. Other advantages of the estimator include ease of construction with high-level, intuitive code (i.e., TensorFlow, Google's open-source machine learning and deep learning library), and unlike models, they do not require high amounts of computational power.

Test Beds

- Mines Park, Golden, CO

Figure 2. Sequencing-batch membrane bioreactor (SB-MBR) located at the Mines Park testbed.

.

I/P Partners

- Aqua Aerobic Systems, Inc., Rockford, IL
- Kennedy/Jenks Consultants San Francisco, CA
- Ramey Environmental, Firestone, CO

Approach

- Data was collected from a demonstration-scale sequencing-batch membrane bioreactor (SB-MBR), which treats 7,000 GPD of municipal wastewater from a student-housing complex in Golden, CO (Figure 2).
- A week's worth of data was gathered on the SB-MBR's conditions, including total suspended solids (TSS levels), dissolved oxygen, temperatures, and ammonium concentrations in the influent, and nitrate concentration in the effluent was collected by a network of "soft" sensors, or software-based sensors.
- An estimator for a DNN that uses regression, a technique used to determine statistical relationships between two or more variables, was developed to predict the nitrate concentration in the effluent water while programmed to use 80% of the data set for training the model and the other 20% for testing the accuracy.
- The DNN was run with multiple data sets, all of which ranged in sizes between 500 data points to 22,000 data points, and then the mean squared error (MSE) of the model was outputted by the program as a measurement of accuracy.

Results

- All monitored SB-MBR variables were normalized to zero mean and unit variance prior to training the estimator. Figure 3 shows the ammonium and nitrate concentrations over the testing period (1 week) after the data was normalized. This graph exemplifies how input data might look post-preparation for training the DNN estimator.

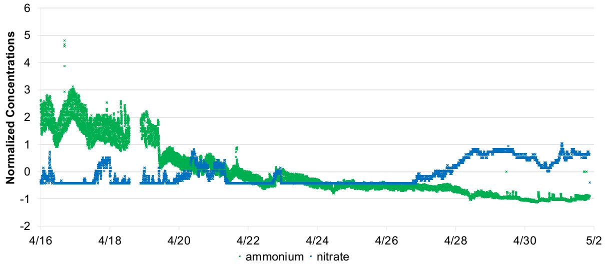


Figure 3. Normalized results of a DNN model to predict SB-MBR effluent nitrate concentrations

- Figure 4 shows where the optimal model fit was achieved based on the amount of data used to train the DNN. Model error declines exponentially below 7,000 data points, which indicates where the model is underfit given the amount of data. The average MSE spikes at 20,000 where the model becomes overfit.

Sections of graph shown in Figure 4:

- 🟦 Model is "underfit" and will not be able to make very accurate predictions
- 🟩 The model has the optimal amount of data for making predictions
- 🟧 The DNN has been fed with too much data or "overfit"

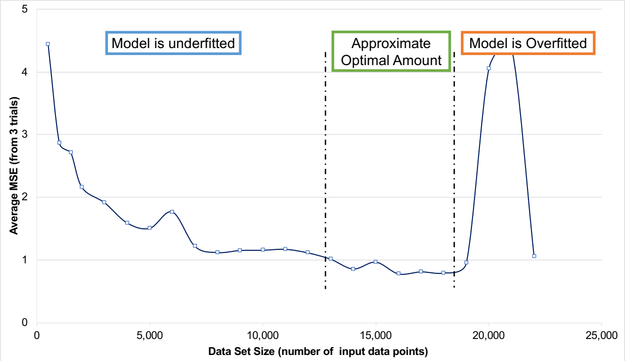


Figure 4. Graphical representation on how the amount of data impacts the accuracy of a DNN

Conclusions

The amount of data used to train a neural network has a rather large impact on the accuracy of the results. There appears to be the expected trend that as the size of the set increased, the accuracy increased up to a threshold. This is the transition from the model being underfitted to more optimized. The spike in MSE indicates where the model becomes overfitted and loses the capability of predicting results with any accuracy. These results indicate that for the data collected over the training period (1 week), the ideal amount of data lies somewhere in the range of 13,000 data points to 18,000 data points.

Given this information, these results can be put to use when applying a neural network to the SB-MBR wastewater treatment system. The outcomes may assist in providing a range for how much data is needed to create an optimized model for an early fault detection system.

Next Steps

Use the results from this research to guide the amount of data used to train or update the neural network used for a fault detection system to predict problems such as biological shifts, flow-interruption, and pump-shutdowns

Acknowledgements

I wanted to thank ReNUWIt for the opportunity to perform research and generously opening up their resources to me.

I also thank TensorFlow for making all of their software open-source and available to use for research projects such as this.

[1] Dzeroski, S, et al. "Classification Of River Water Quality Using Machine Learning." Vol. 5, 1994, doi:10.3897/bdj.4.e7720.figure2f.

Learn more about our research:

www.renuwit.org

Supported by the National Science Foundation at: