ZUSAMMENFASSUNG

Damit die Umweltbelastung durch schadstoffhaltige Produktionsabwässer aus der chemischen Industrie weiter reduziert werden kann, verfolgt die BASF Gruppe den Einsatz von Data Science Methoden. Regelmäßige Laboranalysen von Wasserproben und zahlreiche Sensoren im Abwassernetz liefern umfangreiche Daten und dienen zur Überwachung der Schadstoffemissionen, damit gesetzlich vorgegebene Grenzwerte eingehalten werden.

Im Fokus stehen die mittleren Tageskonzentrationen vom chemischen Sauerstoffbedarf (CSB) und Ammonium-Stickstoff (NH_4-N) sowie das Tagesvolumen an Abwasser. Zur Vorhersage dieser Zielgrößen werden lineare Modelle eingesetzt. Kritische Situationen, wie Grenzwertüberschreitungen, können damit frühzeitig erkannt und vor allem erklärt werden. Als Verbesserung zur Methode der kleinsten Quadrate (KQ-Methode) wird das Lasso-Verfahren bei der Modellkoeffizientenschätzung eingesetzt. Vorteile sind die Vermeidung der Überanpassung und die Erhöhung der Interpretierbarkeit, weil nur einflussreiche Variablen im Modell erhalten bleiben. Der Generalisierungsfehler wird bei NH_4-N um bis zu 90% gegenüber der KQ-Methode reduziert.

Im Praxisbetrieb einer Kläranlage muss mit Ausreißern in den Sensordaten gerechnet werden. Die KQ-Methode und das Lasso-Verfahren reagieren sensibel auf Ausreißer bei der Koeffizientenschätzung, was zur Verzerrung der Modelle führen kann. Um dem entgegenzuwirken, werden auch robuste Schätzverfahren, wie die Quantilregression (QR) für den Median und die Huber-Regression (HR) betrachtet. Der Generalisierungsfehler wird mit der QR bei CSB um 19% reduziert gegenüber dem Lasso-Verfahren.

Sensoren befinden sich an verschiedenen Orten im Abwassernetz und stellen potentielle Prädiktoren dar. Um das Potential eines Prädiktors zur Vorhersage einer Zielgröße voll ausschöpfen zu können, bedarf es einer Korrektur des variablen Zeitversatzes, der aufgrund von Schwankungen im Volumenstrom entsteht, der wiederum die Information eines Prädiktors zum Ort der Zielgröße überträgt. Es wird deshalb ein Verfahren entwickelt, das eine systematische Korrektur dieses variablen Zeitversatz ermöglicht. Bei NH_4-N wird durch dieses Verfahren der Generalisierungsfehlers um 80% reduziert.

Zur Vorhersage des abgegebenen Tagesvolumens gibt es neben dem Volumenstrom keine weiteren informativen Prädiktoren. Der Volumenstrom zeigt autokorreliertes sowie saisonales Verhalten und wird durch ein SARIMA Zeitreihenmodell vorhergesagt. Die Integration des Volumenstromes über der Zeit führt zum Tagesvolumen. Zu Tagesbeginn liegt der mittlere absolute prozentuale Fehler bei 6%. Dieser Wert sinkt bis zum Tagesende auf Null.

Schlagwörter: Lineare Modelle, Lasso, Quantilregression, Huber-Regression, SARIMA, Zeitreihenmodelle, Feature Engineering

ABSTRACT

In order to reduce the environmental impact of pollutant production wastewater from the chemical industry, BASF Group pursues the use of data science methods. Regular laboratory analyzes of water samples as well as numerous sensors, integrated in a wastewater treatment plant, provide extensive data used to monitor pollutant emissions in order to observe legal limit values.

The focus is on the daily mean of the concentrations of CSB and NH4-N as well as the daily volume of wastewater released. Linear models are used to predict these target values. Critical situations, such as limit value exceedances, can be explained and recognized at an early stage. As an improvement to the least squares method (LS), the Lasso method is used to estimate model coefficients. The advantages are avoiding overfitting and increasing interpretability, because only important variables are retained in the model. The generalization error is reduced for NH4-N by up to 90% compared to LS.

In practical operation of a wastewater treatment plant, outliers in the sensor data have to be expected. The LS and the Lasso method react sensitively to outliers during coefficient estimation leading to model bias. Therefore robust methods are also considered such as the QR for the median and the HR. The generalization error is reduced with QR for CSB by 19% compared to the Lasso method.

Sensors are located at different locations in the wastewater treatment plant. All can be considered as potential predictors. In order to be able to fully exploit the potential of a predictor for predicting a target variable, the correction of the variable time offset is needed that arises due to fluctuations in the volume flow, which in turn transmits the information of a predictor to the location of the target variable. Therefore a method is developed to correct this variable time offset systematically. In the case of NH4-N, this procedure reduces the generalization error by 80%.

In order to predict daily volume of wastewater released there are no other informative predictors apart from the volume flow. The volume flow shows autocorrelated as well as seasonal behavior and is predicted by a SARIMA time series model. The integration of the volume flow over time leads to the daily volume. At the beginning of a day, the mean absolute percentage error is 6%. This value drops to zero by the end of the day.

Keywords: linear models, Lasso, quantile regression, Huber-regression, SARIMA, time series models, feature engineering