機械学習による BOD 値を予測する技術の開発

長岡工業高等専門学校 環境都市工学専攻 学生会員 ○ 藤田祐樹
公益社団法人徳島県環境技術センター 非会員 西岡卓馬
長岡工業高等専門学校 環境都市工学科 正会員 川上周司

1. はじめに

合併処理浄化槽 (以下，浄化槽) の維持管理や法定検査は，保守点検業者や清掃業者の現場技術者や指定検査機関の検査員が直接，浄化槽現場に出向いて行っている．しかし，各家庭に点在する膨大な浄化槽を全て回って確認する労力等は多大なものであり，現場における作業時間や労力のみだけでなく，移動に伴う排気ガスによる環境負荷や燃料費など相応の必要経費を要する．また，有機物汚濁を把握する上で重要な水質項目となる BOD 値は，測定期間に 5 日間を要するため，処理状況が思わしくない場合においては，採水時段階で初期対応を行うことは難しい．こうした中でも現場技術者や検査員は，処理水の透視度 (以下，Tr)，曝気槽の DO，スカムの生成状況などから浄化槽の状況を経験的かつ総合的に判断し，対応しているのが現状である [1)]．

本研究では，短時間で BOD 値を測定する技術の開発を目的として，機械学習を用いて浄化槽の BOD 値を予測する技術の開発を行った．まず，対象とする浄化槽の DO 及び Tr と BOD 値を関連付けて学習させるデータセットの作成を行い，機械学習を用いて BOD 値を予測した．

2. 研究手法

2.1 データセットの作成

DO と Tr から BOD 値を機械学習を用いて予測する技術の開発には，まず初めに機械学習モデルに学習させるデータセットの作成を行う必要がある．浄化槽の曝気槽から測定した DO 値と処理水の Tr を説明変数とし，BOD 値を目的変数とした．データセットは合計で二つ作成した．一つは回帰分析に用いた．もう一方は分類分析に用いるため，BOD 値が 20 mg/L よりも大きいデータを 1 とし，BOD 値が 20mg/L 以下のデータを 0 と変換し，二値分類のデータセットを作成した．データセットは，訓練データ，検証データ，テストデータに分け，その数はそれぞれ 3164, 352, 391 とし，訓練データと検証データの分割する割合は 9:1 とした．

2.2 機械学習モデルについて

DO と Tr を関連付けて学習させ BOD 値を機械学習を用いて予測するため，モデルは回帰分析と分類分析の二通りの手段で行った．アルゴリズムは，回帰分析では Extreme Gradient Boosting[2)]を用い，分類分析には Extra Trees Classifier[3)]と Random Forest Classifier[4)]を用いて最適化を試みた．また，ハイパーパラメータの調整に用いた手法はランダムサーチであり，探索回数は回帰分析，分類分析どちらも 500 回とした．

3. 研究結果

3.1 機械学習による回帰分析

回帰分析により得られた学習曲線を図 1 に示す．この学習曲線は，データ数に応じた予測精度を示すグラフである．良好な機械学習モデルの場合，訓練データと検証データの両方に対して同等な予測精度を発揮することから，一般的にデータ数が増加するのに伴い training score が減少し，cross validation score が増加する傾向が見られる．本解析では，データ数の増加と共に，cross validation score は増加しているが training score は減少しておらず，横ばいで推移した．これは，学習不足が生じている際に見られる傾向 [5)]と類似しており，

本解析でも学習不足が考えられた.

次に作成したモデルの予測精度を評価するため，テストデータに対する機械学習モデルの予測値と実測値(正解値) の相関関係を示し，決定係数を算出した (図 2). 結果，決定係数は，0.8255 であり正の相関を示した. しかし，二乗平均平方根誤差 (RMSE) を算出したところ 10.527 と高い値となり，予測精度は低かった. 表 1 に予測値と実測値の誤差を 3 段階に区切った際のデータ数を示す. 予測値と実測値の誤差が±5 mg/L 以内のデータ数は 57.0%しかなく，高い精度を持って予測することはできなかった. RMSE は予測を大きく外すと数値が大きくなるため，テストデータの予測結果で大きな誤差が生じたものが存在していたことが RMSE が高い値となった原因と考えられる.

表 1 誤差が±5，±10，±10 より大きいデータの数とその割合

	誤差範囲		
	±5(mg/L)以内	±10(mg/L)以内	±10(mg/L)より大きいもの
データ数	223	318	72
割合	57.0%	81.3%	18.4%

3.2 機械学習による分類分析結果

回帰分析の結果，学習不足が生じる原因として，回帰分析をするにはデータ数が少なかったと考えた. そこで，回帰分析よりデータ数を必要としない分類分析を行うこととした. しかしながら，BOD 値を 1 mg/L ごとに細かく分け，分類のカテゴリーを増やしていくと同様にデータ数の不足が生じると予想された. そこで浄化槽の排水基準である BOD 値で 20 mg/L を上回るか下回るかという二分類でモデルを構築することとした.

本研究で行った分類分析により得られた学習曲線を図 3 に示す. 結果, 検証データに対する正解率は, 81.9%を示した. 図 3 から，回帰分析と比べ，training score がデータ数の増加とともに減少しており，cross validation score は徐々に上昇している. これらの傾向から学習不足がないと判断し，テストデータを用いて未知のデータに対する予測精度を検証した. その結果，正解率は 77.0%であった.

さらに精度の向上を図るために機械学習のアルゴリズムの変更を行った. 変更後のアルゴリズムは Random Forest Classifier を用い解析を行った. その際の学習曲線を図 4 に示す. 結果，検証データに対する正解率は 82.6%であり，前のモデルよりも上昇した. 図 4 を見ると，学習データ数の増加と共に training score は徐々に減少し，cross validation score は徐々に上昇している. このような傾向から学習不足がないと判断し，次にテストデータを用いて未知のデータ予測精度を検証した. その結果，正解率は 84.1%を示し

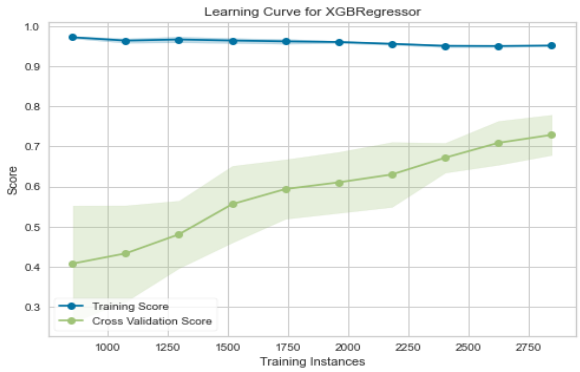

図 1 回帰分析を解いた際の学習曲線. 縦軸は正解率，横軸はデータ数を示している. 青線は訓練データに対するモデルの予測精度を示し，緑線は検証データに対するモデルの予測精度である. Cross Validation Score の曲線に付随して示される緑色の領域は交差検証により求められた制度の最大値と最小値の範囲を示している.

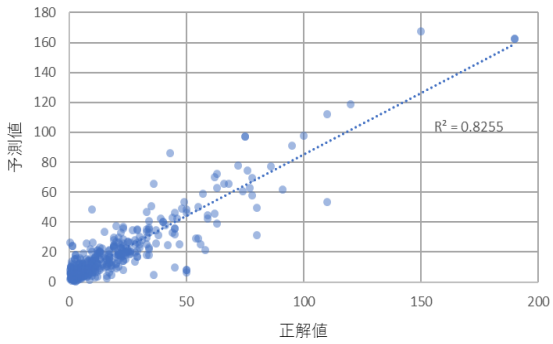

図 2 回帰分析で得られた BOD の予測値と実測値 (正解値) の相関関係. 図中の R^2 値は決定係数を示している.

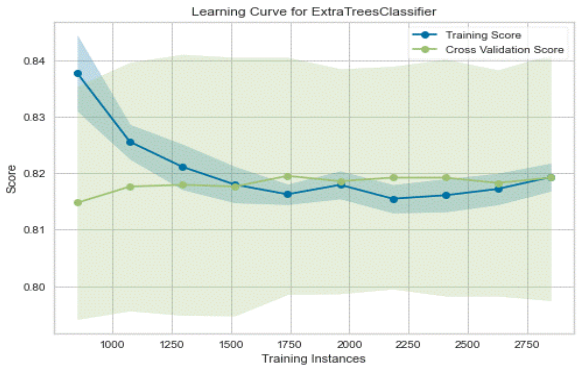

図 2 Extra Trees classifier を用い分類分析を解いた際の学習曲線．縦軸は正解率，横軸はデータ数を示している．青線は訓練データに対するモデルの予測精度を示し，緑線は検証データに対するモデルの予測精度である．Cross Validation Score の曲線に付随して示される緑色の領域は交差検証により求められた制度の最大値と最小値の範囲を示している．

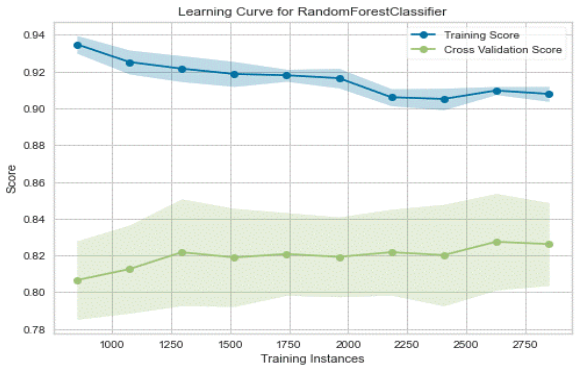

図 3 Random Forest Classifier を用い分類分析を解いた際の学習曲線．縦軸は正解率，横軸はデータ数を示している．青線は訓練データに対するモデルの予測精度を示し，緑線は検証データに対するモデルの予測精度である．Cross Validation Score の曲線に付随して示される緑色の領域は交差検証により求められた制度の最大値と最小値の範囲を示している．

た．アルゴリズムを変更したことにより，正解率が向上した．これら結果から，本研究のデータセットに関しては，Extra trees classifier よりも Random Forest Classifier のモデルの方が最適であると判断した．

3.3　ハイパーパラメータについて

機械学習モデルの学習実行前にハイパーパラメータを調整することで，機械学習モデルの予測精度の向上，過学習の抑制が期待できるため，調整を行った．最適化前と最適化後の正解率の違いを表 2 に示す．また，その際のハイパーパラメータの設定値を表 3，4 に示す．回帰分析は，ハイパーパラメータを調整することで決定係数が約 0.005 上昇したが，分類分析では, Extra trees classifier において正解率が約 4.5%減少し，Random Forest classifier は約 3.9%減少した．これら結果より、本研究ではランダムサーチによるハイパーパラメータの最適化は困難であった．

表 2 ハイパーパラメータを調整することによる評価指標の変化

	決定係数	正解率	
	Extreme Gradient Boosting	Extra trees classifier	Random Forest classifier
最適化前	0.727	86.4%	86.5%
最適化後	0.732	81.9%	82.6%

4. 考察

本研究では DO と Tr のデータから BOD 値を予測する技術を回帰分析，分類分析の二通りの方法で開発した．回帰分析は，BOD 値そのものの値を予測できることから汎用性は高いと思われるが，±5 mg/L の範囲内で予測できたデータ数の割合は約 50%程度と，精度は低かった．それでも予測の範囲を±10 mg/L にまで広げるとその割合は

表 3 Extreme Gradient Boosting のハイパーパラメータ

n_estimators	230
max_depth	7
learning_rate	0.419
verbosity	0
objective	reg:squarederror
booster	gbtree
tree_method	auto
n_jobs	-1
gamma	0
min_child_weight	3
max_delta_step	0
subsample	0.9
colsample_bytree	1
colsample_bylevel	1
colsample_bynode	1
reg_alpha	0.005
lambda(reg_lambda)	10
scale_pos_weight	31.2
base_score	0.5
random_state	123
missing	nan
num_parallel_tree	1
monotone_constraints	()
importance_type	None
gpu_id	-1
validate_parameters	1
predictor	auto
enable_categorical	FALSE
eval_metric	None

80.7%に達した．一方，分類解析は BOD 値が 20 mg/L を超えるか否かを予測する制度では 84.1%を示した．これら解析の予測精度が十分であるか否かについては，議論が必要である．BOD 値は最終的には排水基準を満たすか否かを判断する上で重要な指標である．その上で重要になるのが排水基準となる 20 mg/L という値である．この境界を超えるか否かを判断するのであれば，本研究の結果では分類分析を用いた方が適していたと思われる．本研究では，回帰分析においてデータ数の不足が示唆された．本研究で集積したデータを見ると，BOD が 20 mg/L 以下という良好な処理水が得られた際のデータ数が多く，偏っていることがわかる (図 2)．一般的に浄化槽は安定的な処理能力を発揮することから，そもそも処理水の悪化したデータ数が少ないという問題がある．このような偏ったデータでモデルを構築すると，十分な予測精度が出ないという報告もある．したがって，BOD 値が均等かつ十分なデータが準備しづらい現状においては，分類分析を用いる方が実用性が高いと考えている．

表 4 Extra trees classifier と Random forest classifier のハイパーパラメータ

	Extra Trees classifier	random forest classifier
n_estimators	60	180
criterion	gini	entropy
max_depth	8	11
min_samples_split	5	5
min_samples_leaf	3	2
min_weight_fraction_leaf	0.0	0.0
max_features	1	sqrt
max_leaf_nodes	None	None
min_impurity_decrease	0.001	0.0001
bootstrap	FALSE	FALSE
oob_score	FALSE	FALSE
n_jobs	-1	-1
random_state	123	123
verbose	0	0
warm_start	FALSE	FALSE
class_weight	None	None

5. まとめと今後の展望

本研究により，DO と Tr から処理水の BOD を予測するモデルを構築できた．予測精度の面で課題は残るが，本手法を用いることで浄化槽の処理状況が思わしくない場合に，少なくとも従来の BOD 値の測定に必要である日数の 5 日よりも早く対応をとることが可能になると思われる．

本研究で用いたハイパーパラメータの調整法では最適なパラメータが見つからなかったことから，ランダムサーチ以外の調整法であるベイズ最適化，変数増減法などを試す予定である．また，今後はさらなる精度向上に向けてニューラルネットワークを用いた解析手法なども検討していく予定である．

6. 参考文献

1) 中島進, 大町盛一郎, 処理水の性状に着目した水質悪化施設の原因究明フローの構築と早期改善への取り組みについて, *全国浄化槽技術研究集会資料*, 2018.
2) Chen, T., Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016.
3) Geurts, P., Damien Ernst, Louis Wehenkel. Extremely randomized trees, *Machine Learning*, Vol. 63, pp.3–42, 2006.
4) Breiman, L., Random Forests. *Machine Learning*, Vol. 45, pp.5–32, 2001.
5) Raschka, S., Vahid Mirjalili. Python Machine Learning, Second Edition, Chapter 1. Packt Publishing. 2017.