オンサイトで取得できるデータを用いた深層学習による浄化槽の BOD 予測技術の開発

長岡工業高等専門学校 環境都市工学専攻科　非会員 ○ 久住葉瑠
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. 背景と目的

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 (Tr)，曝気槽の溶存酸素 (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，現状 5 日間かかる BOD 検査の時間短縮を目的として，深層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．

2. 研究手法

2.1 画像データの加工

対象とした浄化槽は, 岩手県と埼玉県の各地のものを使用した. 今回の解析では, 水質データの DO と透視度 (Tr) を使用するため岩手県のみの浄化槽の画像を使用した. 画像を浄化槽部分のみを抽出することを目的にして. トリミングを行った (図 1). さらに, 水質データは標準化を行った.

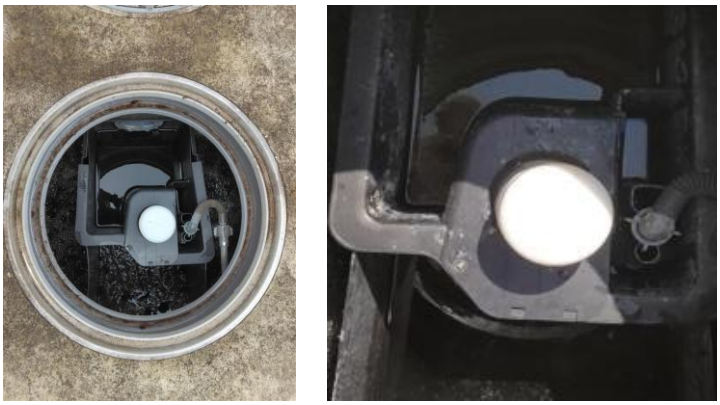


図 1　元の画像とトリミング後の画像

2.2 データセットの作成

画像と水質データを関連付けて BOD 値を予測する技術の開発には,まず初めに深層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．浄化槽の好気槽の上部写真, 水面の Tr と DO を説明変数, BOD を目的変数とした．そのデータセットの BOD 値を分類分析に用いるため, BOD 値が 20 mg/L 以上のデータを 1 とし, BOD 値が 20 mg/L 未満のデータを 0 と変換し, データセットを作成した．データセットは, データを学習させるための訓練データとモデルの確認やパラメータを調整するための検証データ, 学習済みモデルの汎用性を評価するためのテストデータに分割する. 本研究では収集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, 偏っていた．本研究では, 二値のデータ数を近づけるため, 20 mg/L 以下のデータを削減するアンダーサンプリングを行った.

2.3 深層学習モデルについて

本研究では, 畳み込みニューラルネットワーク (Convolutional Neural Network) 構造を用いてモデルを構築した．転移学習モデルを使用した解析には resnet18 [2]を採用し, ファインチューニングは, ImageNet [3]と呼ばれる 1,400 万枚以上の画像の教師ラベル付き画像データベースを用いて事前学習した resnet18 に収集した浄化槽の画像を用いて再学習させた．モデルの学習, 推定には k 分割交差検証法とホールドアウト法を用いた.

本研究では, Python を基に構築されているニューラルネットワークライブラリの PyTorch を用いてニューラルネットワークの構築,学習及び評価を行った.　損失関数はバイナリクロスエントロピーを用いた.

2.4 正解率と適合率と再現率

本研究では, 予測した BOD 値が正解の BOD 値とどの程度一致しているかの正解率 (Accuracy) と, 学習モデルが正と予測したものが, 本当に正しかった割合の再現率 (Recall) , 実際に正のものをどれだけ正しく予測できたかの適合率 (Precision) を計算し, 最適なモデルに向けた改善を行った.

2.3 学習曲線

学習曲線は, AI モデルの学習過程を可視化したものであり, 横軸に学習回数, 縦軸に損失値または予測精度をとって示した. 損失値の学習曲線は, AI モデルが訓練と検証データに対して誤差をどの程度生じているかを示す指標であり, 学習が進むにつれ 損失値が減少する傾向にある場合は, AI モデルが訓練データから特徴を学習していると判断した. 反対に, 損失値が上昇する傾向がある場合は, AI モデルが訓練データに過度に適合している状態 (過学習) だと判断した. 予測精度の学習曲線は, AI モデルがどの程度正確に予測を行っているかを示す指標であり, 学習が進むにつれて予測精度が上昇する傾向にある場合は, AI モデルが訓練データから特徴を学習していると判断した. 反対に予測精度が減少する傾向にある場合は, 過学習が発生していると判断した.

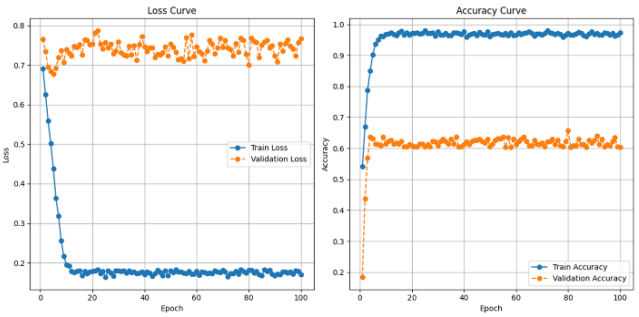
2.5 混同行列

モデルの評価指標として混同行列を用い,予測結果と実際のクラスとの関係を可視化した. 混同行列は縦軸に正解ラベル, 横軸に予測ラベルを配置した行列であり, True Positive, False Positive, True Negative, False Negative の内訳を確認できる. 本研究では,各クラスの分類性能や誤分類パターンを把握するために活用し, モデルの改善に役立てた.

3. 研究結果

3.1　ホールドアウト法の初期検討 (解析 1)

データセットは,それぞれ訓練データが 788 個, 検証データが 353 個, テストデータが 276 個とした. 得られた学習曲線と混同行列を下に示す (図 2). テストデータへの予測では, Accuracy 60.9% , Precision 46.4% , Recall 45.5% となった. 得られた学習曲線とテストデータへの予測から汎化性能があがっていないことが分かる. 混合行列を見ると, 20 mg/L 未満と予測して正解しているものが 1 番多かったが, 20 mg/L 以上と予測し正解したものが 1 番少なかった. 学習曲線を見ると検証デーの loss が途中から横ばいなっていたことから過学習という状況と考えられる. Dropout や L2 正則化などのパラメータ調整を行うことで過学習が抑えられると考えた.

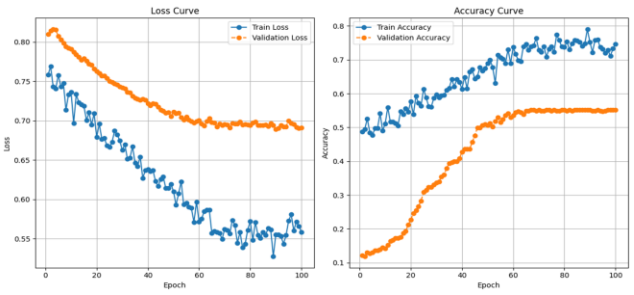


		Predicted Labels	
		0	1
True Label	0	123	52
	1	56	45

図 2　解析 1 の学習曲線と混同行列

3.2　パラメータ調整 (解析 2)

過学習対策で Dropout や L2 正則化のパラメータ調節を行った. 得られた学習曲線と混同行列を下に示す (図 3). テストデータへの予測は, Accuracy 55.5% , Precision 43.0% , Recall 46.0% となった. テストデータと学習曲線から学習不足という状態なことが分かった. これは Dropout や L2 正則化などのパラメータを強めすぎたことで, モデルが複雑なパターンを学習できなかったと考えた. パラメータの変更では, 改善が見込めないと考え, 最適化アルゴリズムと学習率スケジューラの変更, データ拡張を行った.

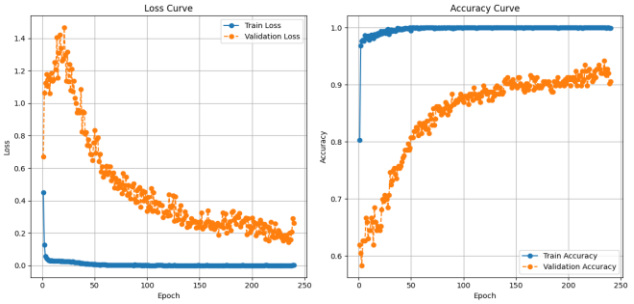


		Predicted Labels	
		0	1
True Label	0	123	55
	1	52	46

図 3　解析 2 の学習曲線と混同行列

3.2　最適化アルゴリズムと学習率スケジューラ変更 (解析 3)

訓練データを拡張し 1141 個とした. そして, 最適化アルゴリズムを, Adam を用いて学習を行っていたが, Weight Decay 項を勾配更新から切り離して扱うことで, 過学習の抑制と学習の安定性向上が期待できる AdamW へ切り替えを行った. さらに, 学習率スケジューラを ReduceLROnPlateau から, 学習率を余弦関数に沿って周期的に減衰させることで,局所解への収束を回避しながら,効率的にグローバル最適解を探索できると報告されている CosineAnnealingLR に変更した.



		Predicted Label	
		0	1
True Label	0	233	58
	1	48	152

図 4　解析 3 の学習曲線，混同行列

CosineAnnealingLR は epochs が 20 進むごとに学習率を減少させ, 1e-6 を最小値として設定した. 得られた学習曲線と混同行列を下に示す (図 4). テストデータへの予測では, Accuracy 78.4%, Precision 72.4%, Recall 76.0% となった. 学習曲線から, 検証データの loss が学習初期は上がるが, 徐々に減少した. 検証データの Accuracy も 90% 程度まで増加したため, モデルの学習が行えたと判断した. 閾値を変更することで予測精度は 80%を超えた. しかし, 20 mg/L 以上への予測が下がってしまうため, 実際に現場で使用することを想定したら閾値の変更は行わないほうがよいと考えられる. 解析でモデルが予測を間違えた画像を示す(図 5). 多くの間違いは, パイプやふたなどのノイズが原因であることが分かったが図 8 右図のように水面が見えているものも間違えることがあった. 水面が見えていても暗いとモデルが予測を迷ってしまっている傾向にあった.

図 5　AI モデルが予測を間違えた画像

4. 考察

本研究では画像データと水質データから BOD を予測する技術を分類分析で開発した. 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[4]という値である. 本研究では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った．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，少ない方のデータを拡張するオーバーサンプリングか，多い方のデータ数を削減するアンダーサンプリングを行う必要であると考えた．

モデルが間違えてしまった画像の中には，水面が見えているが間違えてしまっているものもあった．画像で判断しにくいものは，水質データからの予測を行えるように，水質データの重みも検討していくことでモデルの精度が向上すると考える．

5. まとめと今後の展望

本研究では，浄化槽の画像と取得可能な水質データを用いた BOD 予測のモデル開発を試みた．今回の解析では,ホールドアウト法を用いて BOD を 80%程度予測するモデルの構築を行うことができた．本研究ではモデルの構築に向け，各種パラメータ,や最適化アルゴリズム，学習率スケジューラを変更した．特に，最適化アルゴリズムの AdamW および学習率スケジューラの CosineAnnealingLR が汎化能力の向上に有効であることが確認された (表 1)．水面が見えている画像の予測を間違えてしまうという課題も出た．今後は，モデルの精度向上として，さらに詳細なパラメータ調整,やデータ数の増加，水質データの重みの検討を行っていき予測精度の高い信頼性のある学習モデルを作成することで，現状 5 日間必要とされている BOD 値の測定を，今よりも迅速に推測できる技術の確立そして実装を目指す．

表 1　解析のまとめ

	Drop out	weight decay	optimizer	lr_scheduler	損失関数	モデルの構築
解析1	0	0	Adam	ReduceLROnPlateau	BCEWithLogitsLoss	×

	Drop out	weight decay	optimizer	lr_scheduler	損失関数	モデルの構築
解析2	0.3	0.4	Adam	ReduceLROnPlateau	BCEWithLogitsLoss	×

	Drop out	weight decay	optimizer	lr_scheduler	損失関数	モデルの構築
解析3	0.3,0.4	0.4	AdamW	CosineAnnealingLR	BCEWithLogitsLoss	○

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) He, Kaiming, et al. "Deep Residual Learning for Image Recognition.". arXiv:1512.03385, 2015
3) Deng, J., Dong, W. *et al.* "ImageNet: A Large-Scale Hierarchical Image Database", *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2009.
4) よりよい水環境のための浄化槽自己管理マニュアル - 環境省, 2009