

模擬異常データを用いた深層学習による従属栄養細菌数迅速計測手法の高精度化

長岡工業高等専門学校 非会員 ○星惇稀, 徳嵩友志, 正会員 川上周司

1. はじめに

従属栄養細菌は、上水試験法では「有機物を比較的低濃度に含む培地を用いて低温で長時間培養したとき、培地に集落を形成するすべての細菌」¹⁾と定義され、水道施設の清浄度の劣化を端的に表す指標として、水質管理に有用とされている。しかし、R2A 寒天培地を用いて「温度を 19~21℃に保持できる低温恒温器内で 7 日間培養」²⁾する標準的な培養法では、細菌数の測定結果が得られるまでに 7 日を要する。国内の給配水に要する時間は培養に必要な日数よりも短いため、結果を水質管理にフィードバックすることは難しい。蛍光染色を用いる、培養に依らない検出法もあるが、高価な蛍光顕微鏡や高度な専門知識が必要で、浄水場での簡便な運用には向かない。そこで本研究では、蛍光顕微鏡を必要としないグラム染色の顕微鏡画像と深層学習を組み合わせた迅速計測を目指している。

しかし、浄水場は本来清浄な水を供給する施設である。そのため、深層学習に必要な異常データとなる、細菌数が基準値を超えた試料の画像が十分に得られない。また、実際の運用では、河川水や下水の混入、凝集剤不足など原因の異なる多様な異常が起こりうるため、限られた異常のみを学習したモデルでは、未知の異常に対応できない可能性がある。そこで本研究では、これらの異常状態を模した試料を作成し、その画像を模擬異常データとして学習させることにより、多様な異常に対して精度を保てる汎用性の高いモデルの構築を最終目標とした。

本報告では、模擬異常データを組み込む前段階の基礎的な検討として、細菌数が基準値を超えるかどうかを正常・異常に分類するモデルを構築した。1 枚の画像を入力した場合と、同一試料の 10 枚の画像を 1 つのセットとして入力した場合とで予測精度を比較し、入力画像の枚数が予測精度に与える影響を明らかにすることを目的とした。

2. 方法

2.1 データセットの作成

本研究では、長岡市水道局妙見浄水場の原水、沈殿処理水、浄水を採水した。各試料 100 mL を孔径 0.2 μm のメンブレンフィルターで濾過し、フィルター上でグラム染色を行ったものを、位相差顕微鏡を用いて倍率 400 倍で観察し、各試料から約 100 枚の画像を取得した。同じ試料を R2A 寒天培地で $20\pm 1^\circ\text{C}$, 7 日間培養し、得られた細菌数を教師データとした。従属栄養細菌は水質管理目標設定項目の一つであり、「1 mL の検水で形成される集落数が 2,000 以下(暫定)」³⁾と定められている。そのため、本研究では 2,000 CFU/mL を閾値とし、超える試料を異常、超えない試料を正常と定義した。浄水場には異常データが本来存在しないため、本検討では暫定的に、細菌数の多い原水・沈殿水の試料を異常データの代用として用いた。画像は 2048×2048 にクロップし、256×256 にリサイズした。

本研究では、入力画像の枚数が予測精度に与える影響を明らかにするため、Set1 を画像 1 枚のみを入力する条件、Set2 を画像 10 枚を 1 つのセットとして入力する条件として比較した。Set1 と Set2 の比較にあたり、いずれの条件でも訓練データと検証データは 8:2 とし、同一試料の画像が両方に混在しないよう試料単位で分割した。

2.2 分類モデル

Set1 と Set2 では、特徴量の抽出に共通して 5 層の畳み込みニューラルネットワーク (Convolutional Neural

Network: CNN) を用い、画像の融合の有無のみを変えて比較した。CNN は、線や形などの局所的な特徴を調べ、それを段階的に組み合わせて、画像全体の特徴を捉える手法であり、画像認識に広く用いられている。Set1 では、1 枚の画像を CNN に入力し、得られた特徴量から直接、正常・異常を分類した。Set2 では、複数インスタンス学習 (Multiple Instance Learning : MIL) を用いた。MIL は、複数の画像をまとめた。1 つのセットに対して、1 つの正解(正常・異常)を与えて学習する手法である。本研究では、セット内の各画像から CNN で特徴量を抽出し、それらを融合して、セット全体で正常・異常を分類した。融合には Gated Attention を用いた。Gated Attention はセット内の各画像に学習によって重みを付与する手法で、細菌を含む画像を重視して判断できる。評価指標は Accuracy (正解率) と loss (学習損失) とした。

3. 結果および考察

Set1 の学習曲線を図 1 に、Set2 の学習曲線を図 2 に示す。検証データの Accuracy, loss は、Set1 が 0.518, 3.961, Set2 が 0.827, 0.284 であった。Set1 では、学習開始直後から過学習が進んだ。同一試料内でも画像ごとに細菌分布の偏りが大きいことが原因と考えられる。一方、Set2 では、Set1 と比較して Accuracy が大きく向上した。細菌が写っていない画像があっても、複数枚をまとめて判断でき、Gated Attention が細菌を含む画像を重視したため、精度が向上したと考えられた。以上の Set1 と Set2 の比較から、本研究においては、複数枚の画像を一組として入力する手法が有効であることが示された。

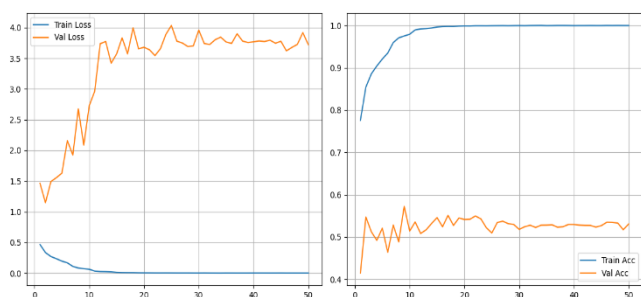


図 1 Set1 の学習曲線。

左図 : loss, 右図 : accuracy.

青線 : 訓練データ, オレンジ線 : 検証データ。

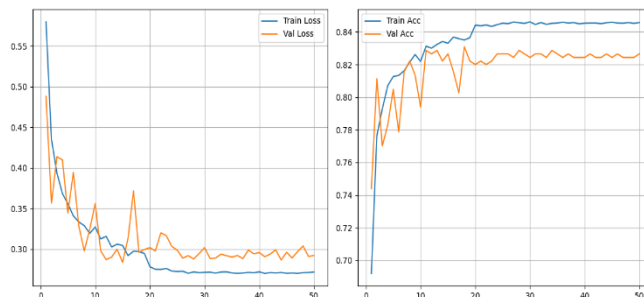


図 2 Set2 の学習曲線。

左図 : loss, 右図 : accuracy.

青線 : 訓練データ, オレンジ線 : 検証データ。

4. 現状と今後の課題

本研究の分類モデルでは、浄水過程の原水・沈殿水を異常データの代用とした。そのため、実際の運用で想定される下水の混入、凝集剤不足などによる細菌数の急増に対しては、学習データの範囲を超えた予測となり、精度が低下する可能性がある。そこで、今後の研究ではこれらの異常状態を模した試料を作成し、模擬異常データとして学習に加えることで、実際の運用に近い条件での精度向上を目指す。

参考文献

- 1) 日本水道協会 (2020), 上水試験方法, IV. 微生物編.
- 2) 厚生労働省 (2007), 水質基準に関する省令の一部改正等における留意事項について, 健水発第 1115002 号.
- 3) 厚生労働省 (2007), 水質基準に関する省令の一部改正等について(施行通知), 健発第 1115003 号.