

ディープラーニングを用いた河川の定点画像から 水質を推測するシステムの開発

阿南高専 賛助会員 ○泉良樹 阿南高専 賛助会員 遠野竜翁 阿南高専 非会員 岡本浩行
阿南高専 非会員 大田直友 長岡高専 賛助会員 川上周司

1. はじめに

水質調査の目的は、環境基準に定められた水質環境基準を維持、達成されているかどうかを監視することである。現在、行われている水質調査は、対象とする河川の現場に向かい、採水を行った後、種々の測定機器を用いて水質の測定を行っている。しかし、水質測定項目には、現場で測定を行える項目とサンプルを持ち帰り実験室で測定を行う項目があり、水質測定が現場で完結できないという問題がある。また、環境省が定期的に水質の測定を行い、周辺状況の観察や水質の把握を行っているが、環境省が定めた地点と時期のみでしか調査が行われていない。そのため、河川の連続的な水質の変化を把握することは現状では難しい。

本研究では、現在考えられる2点の問題点を改善するため、水質データと河川の写真を用いたディープラーニングによる河川の水質を推測するシステムの開発とその実現性の検証を行った。また、水質推測システムの精度が最も高い際の最適な教師データのパターンについて検討した。

2. 実験方法

2.1 教師データの作成

水質を推測するシステムを作成するためには、システムに学習させる教師データを作成する必要がある。教師データは、橋梁から河川の中央部を撮影した定点画像とその時に採取した河川水の BOD を測定し作成した。河川は、徳島県阿南市に流れている河川を対象とした。過去の水質データ¹⁾を参考にし、BOD の値を基準として桑野川（一級河川）と打樋川（二級河川）を選定した。

2.2 水質推測システムの作成

本研究で開発した水質推測システムは、教師あり学習の畳み込みニューラルネットワークを適用し、水質の状態を推測した。構築した学習モデルは、入力層、4 層の畳み込み層とプーリング層、全結合層、出力層で構成される。水質の状態の推測には分類のアルゴリズムを採用し、教師データを訓練データとテストデータに分けて学習モデルで解析を行い、BOD の値を推測する正解率の精度を求めることで水質推測システムの評価を行った。

2.3 教師データのパターンについて

本研究では、教師データの BOD 値やデータ数の違いが構築したシステムの正答率にどのような影響を及ぼすかを検討するために教師データ数を複数のパターンで準備し作成した。具体的には、無調整（パターン1）、教師データ数が BOD 値 0~10mg/L までで均一にある場合（パターン2）と、また分類を BOD 値で 1mg/L 刻みで行う場合（クラス

表-1 パターン 1, 2 の教師データ数

| 解析方法 | 教師データのサンプル数 | | | | | | | | | | 計 |
|-------|-------------|------|------|-----|-----|-----|-----|-----|-----|-----|------|
| | クラス分けなし | | | | | | | | | | |
| | BOD (mg/L) | | | | | | | | | | |
| | 0.5 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | 6.0 | 7.0 | 8.0 | 9.0 | |
| パターン1 | 309 | 1169 | 1130 | 293 | 132 | 59 | 172 | 127 | 165 | 49 | 3065 |
| パターン2 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1000 |

表-2 パターン 3, 4 の教師データ数

| 解析方法 | 教師データのサンプル数 | | | | | | | | | | | 計 |
|-------|-------------|-----|-----|-----|-----|------|------------|-----|------|------------|------|------|
| | 6クラス | | | | | | 3クラス | | | 2クラス | | |
| | BOD (mg/L) | | | | | | BOD (mg/L) | | | BOD (mg/L) | | |
| | 0~1 | 1~2 | 2~3 | 3~5 | 5~8 | 8~10 | 0~2 | 2~5 | 5~10 | 0~3 | 3~10 | |
| パターン3 | 200 | 200 | 200 | 200 | 200 | 200 | 400 | 400 | 400 | 600 | 600 | 1200 |
| パターン4 | 400 | 400 | 400 | 400 | 400 | 400 | 800 | 800 | 800 | 1200 | 1200 | 2400 |

分けなし) や環境基準の類型で分ける場合 (6 クラス)、ある程度幅を持たせた場合 (3 クラス、2 クラス) といった、複数パターンを作成した (表-1, 2)。学習モデルの学習回数は、100 回、300 回、500 回で検証を行った。

3. 実験結果及び考察

3.1 BOD 値ごとに教師データ数を均一化する必要性について

表-3 に各パターンで構築した水質予測システムの学習回数ごとの正答率を示す。教師データを均一化しなかったパターン1の正解率は、訓練データ約78~90%、テストデータ約68~80%と高い正解率を示した（表-3）。しかし、教師データで使用しなかった河川の画像を使って実際に水質の推測を行うと正答できないケースのほとんどが BOD 値が高い場合に偏っていた（データ非表示）。これは、水質データを取得した通りに教師データを作成すると河川の特性上どうしても BOD 値が低い時のデータ数が多くなり（表-1）、学習を進める中で BOD 値が低いものには過学習が起き、BOD 値が高くなると情報不足になっていることが原因と思われる。このシステムの実用化を考えた場合、河川の BOD 値が高くなった際を瞬時に検出したいというニーズがあると考えると、このシステムでは不十分であると判断した。そこで、教師データ数を BOD 値ごとに均一化したパターン2を作成し、システムの改善を試みた。結果、全体の正答率は若干低下したが、パターン1で見られたような BOD 値の高低による差は見られなくなった（データ非表示）。このことから様々な BOD 値に対応できるシステムを作成するためには、BOD 値ごとに教師データ数を均一化することが有効だと考えられた。

表-3 解析方法ごとの正解率

| 解析方法 | クラス分け | 学習回数ごとの正解率 (%) | | | | | |
|-------|---------|----------------|-----|------|-----|------|-----|
| | | 100回 | | 300回 | | 500回 | |
| | | 訓練 | テスト | 訓練 | テスト | 訓練 | テスト |
| パターン1 | クラス分けなし | 78 | 68 | 90 | 78 | 90 | 80 |
| パターン2 | クラス分けなし | 80 | 68 | 90 | 78 | 90 | 70 |
| パターン3 | 6クラス | 70 | 52 | 90 | 70 | 90 | 60 |
| | 3クラス | 50 | 40 | 85 | 70 | 95 | 80 |
| | 2クラス | 85 | 75 | 95 | 80 | 98 | 78 |
| パターン4 | 6クラス | 77 | 67 | 90 | 78 | 90 | 70 |
| | 3クラス | 90 | 70 | 93 | 83 | 95 | 85 |
| | 2クラス | 90 | 80 | 95 | 80 | 98 | 85 |

そこで、教師データ数を BOD 値ごとに均一化したパターン2を作成し、システムの改善を試みた。結果、全体の正答率は若干低下したが、パターン1で見られたような BOD 値の高低による差は見られなくなった（データ非表示）。このことから様々な BOD 値に対応できるシステムを作成するためには、BOD 値ごとに教師データ数を均一化することが有効だと考えられた。

3.2 推測する BOD 値に幅を持たせた際の正答率について

BOD 値を 1mg/L 刻みで推測することは難しいと考え、推測値に幅をもたせたパターンを作成し、正解率が向上できないか検討した。BOD 値を 1mg/L 刻みで分類したクラス分けなしや環境基準の類型に合わせた6クラスのように推測する分類数が細かい時と、2クラスや3クラスのように推測する分類数が粗い時を比較すると推測する分類数が粗い時の方が全体的に精度が上昇する傾向が見られた（図-1, 2, 3）。さらに2クラスと3クラスで比較すると2クラスの方が精度が高かった（表-3）。このことから、推測する分類数を減らすほど精度が上がり、本研究では2クラスに分けた分類が最も精度を上げることが出来ると示唆された。

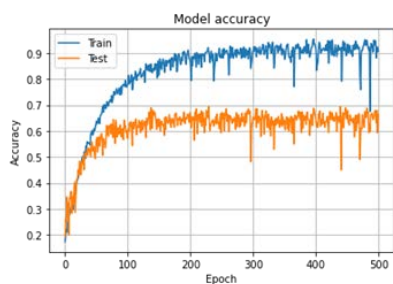


図-1 6クラスの時の正解率

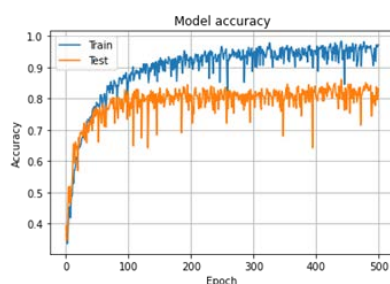


図-2 3クラスの時の正解率

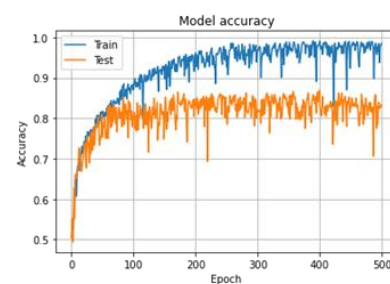


図-3 2クラスの時の正解率

4. まとめ

本研究で開発した河川の画像から BOD 値を推測するシステムは、教師データの数を分類 BOD 値ごとに均一化することで BOD 値を 1mg/L 刻みの精度で予測でき、正解率 78%を達成できた（パターン2、学習回数300回）。また推測値に幅を持たせることでその正率は 85%程度（パターン3, 4、学習回数500回）まで上昇した。このことは本システムが簡易的ではあるが河川の汚濁が進んでいるかどうかを判断できるだけの精度を有していたことを示している。しかし、訓練データとテストデータの正解率に約10%~20%程度の誤差があることも示された。これは過学習の可能性が考えられ、本システムの更なる修正および改良をする必要とするものである。今後はシステムの正則化²⁾やドロップアウトなどの手法を活用することで訓練データとテストデータの誤差を小さくし過学習の解消を目指していく予定である。

5. 参考文献

- 1) 阿南市の環境, 阿南市, 2015
- 2) 鈴木大慈, 過学習と正則化, 応用数理, Vol.28 No.2 Page.76-81, 2018.