3K-03

SVM によるリアルタイム河川水中大腸菌濃度予測

小林　美里 †　　佐野　大輔 ‡　　加藤　毅 †

† 群馬大学理工学部，‡ 北海道大学工学研究院

1　はじめに

水利用における微生物学的安全性は，大腸菌等に代表される「指標微生物」の水中濃度を監視・制御することにより担保されてきた．この「指標微生物」を用いた各種水質基準の運用により，様々な病原菌に起因する水系感染症の発生が抑制されていることは疑いようのない事実である．かつて江戸末期から明治初期に年間１０万人超の死者を度々出したコレラ等による水系感染症の発生件数が，当時とは比較にならないほど激減していることはその証左と言えよう．その一方で，水域における微生物学的水質異常のリアルタイムモニタリングは，未だ実現していない重要課題である．そもそも「指標微生物」によって微生物汚染を「リアルタイムに監視」することは想定されていない．仮に近年急速に発展している分子生物学的な手法を水中病原体のモニタリング手法として取り入れたとしても，水中病原体の存在を検知するのは早くて数時間後であり，リアルタイムモニタリングを実現する際には明白な技術的障壁が存在する．しかも浄水場等において分子生物学的手法を用いて継続的に水中病原体濃度をモニタリングするには多大な経費とマンパワーが必要であり，昨今の税収減に苦しむ行政の立場からすれば，水処理の現場において分子生物学的手法による水中病原体のリアルタイムモニタリングシステムを導入することは，将来的な技術革新に期待を掛けたとしても，実行に移す段階には残念ながら達していないのが現状である．

本研究では，これまでの筆者らの水中病原体のための情報科学的解析の経験 [1] を活かし，分子生物学的手法による水中病原体のリアルタイムモニタリングを行う代わりに，リアルタイムで入手できる水文水質データを使って，病原体濃度を情報科学的に予測するアプローチを採用する．この目標の最初のステップとして，大腸菌濃度にターゲットを絞り，機械学習技術を活用してどの程度予測できるか計算機実験を行った．本稿では，その実験結果を報告する．

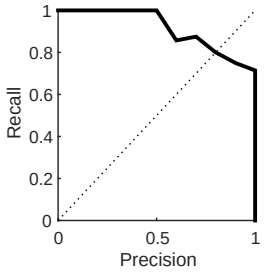

図 1　**Precision-Recall 曲線．**$(\lambda, m) = (0.1, 0.1)$

2　説明変量と目的変量

機械学習技術に入力するための，リアルタイムで入手できる情報として，水文水質データを用いる．具体的には，次の説明変量を用いる．

- WT(℃)：水温．水温が高いと大腸菌濃度が増加する可能性がある．
- pH：水素イオン指数．水素イオン指数が中性付近（7.0–7.5）より高いもしくは低いと大腸菌の生育速度が下がり，異常に高いもしくは低い場合には死滅する可能性がある．
- DO(mg/L)：溶存酸素．水中に含まれる酸素量が高いと大腸菌の増殖に有利である．
- SS(mg/L)：浮遊物質・懸濁物質．粒径が 1mm〜2mm の粒子状物質の含有量を表す．この値が高いと，大腸菌が紫外線等からの不活化ストレスから防御され，生存に有利となる可能性がある．
- BOD(mg/L)：生物化学的酸素要求量．好気性微生物によって有機物が分解されるときに消費される酸素の量を表す．この値が高い場合、大腸菌の濃度も高い可能性がある．
- COD 酸性法 (mg/L)：酸性高温過マンガン酸法の測定値．この値が高い場合，大腸菌の濃度も高い可能性がある．
- COD アルカリ法 (mg/L)：アルカリ性高温過マンガン酸法．この値が高い場合，大腸菌の濃度も高い可能性がある．
- T-N(mg/L)：全窒素．窒素量が高いと大腸菌の生

育に有利となる可能性がある．

- T-P(mg/L)：全リン．リンが多いと大腸菌の生育に有利となる可能性がある．
- 流量 (m^3/s)．流量が多いと大腸菌が希釈され，濃度が低下する可能性がある．
- Station．7 か所の Station で大腸菌を測定した．これを 7 次元のダミー変数表現であらわした．

目的変量は大腸菌濃度である．500 MPN/100ml を仮の基準値と見立てて，基準値以上を陽性，基準値未満を陰性とした 2 クラス分類問題とした．データの収集は，2011 年 12 月 5 日〜2013 年 4 月 17 日の期間で行った．2011 年 12 月 5 日〜2 月 21 日までのデータを訓練用に用い，2011 年 3 月 7 日以降のデータを評価用に用いた．訓練用データの陽性数及び陰性数は，それぞれ，85 個, 83 個で，評価用データの陽性数及び陰性数は，それぞれ，10 個, 6 個であった．説明変量のうち，ダミー変数以外の各変数は，訓練用データの中で値の二乗の平均が 1 になるように，スケーリングを行った．欠損値はその説明変量の平均で，測定値限界未満は 0 で埋めた．

3 サポートベクトルマシン (SVM)

本研究では，説明変量から大腸菌濃度が基準値未満か，基準値以上か予測するタスクに焦点を当て，2 クラス識別器を用いる．線形識別器を採用する．線形識別器の識別関数は，説明変量 $\boldsymbol{x} \in \mathbb{R}^d$ に対して，$\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ で与えられる．その識別関数にある閾値を定めることによって，陽性か陰性か分類する．識別関数に含まれる $\boldsymbol{w}$ は，サイズ n の訓練用データセット $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ $(i = 1, \ldots, n)$ から SVM 学習によって得る．SVM 学習とは，目的関数

$$f(\boldsymbol{w}) := \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} l(y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)$$

の最小解をもって $\boldsymbol{w}$ の値を決定する方法である．ただし，λ は正則化パラメータと呼ばれる正の定数である．$l(\cdot)$ は損失関数である．本研究では，**スムース化ヒンジ損失** (smoothed hinge loss)

$$l(z) := \begin{cases} 1 - z - m/2 & \text{if } z \in (-\infty, 1-m], \\ \frac{1}{2m}(z-1)^2 & \text{if } z \in (1-m, 1], \\ 0 & \text{if } z \in (1, +\infty) \end{cases}$$

を採用した．通常よく用いられてきた標準的なヒンジ損失はスムース化ヒンジ損失において $m \to 0$ としたものに等しい．スムース化ヒンジ損失は $(1/m)$-スムース，すなわち，導関数が $(1/m)$-リプシッツ連続である．スムースではない損失関数を用いると，学習に用いる目的関数 $f(\boldsymbol{w})$ もスムースではなくなり，一般に多くの確率的勾配法において最適解への収束が遅くなる．標準的なヒンジ損失を実装したフリーの SVM のソフトウェアはいくつか利用可能だが，それらの多くは，しばしば最適解への収束を待たずに計算を停止してしまっている．筆者らは，もし最適解に十分近くない解を使って予測してしまうとすると，ソフトウェアごとに異なる解になり，予測結果の一意性と再現性が損なわれ，科学的検証に支障がでるとの見地に立ち，最適解に到達可能な損失関数として，スムース化ヒンジ損失を採用するに至った．最適化のために Stochastic DCA 法を Matlab で実装した．双対性ギャップ 10^{-6} 以下を停止条件として，最適化アルゴリズムを動かした．

4 計算機実験

識別関数の閾値 1 つにつき，Precision と Recall のペアが 1 つ得られ，図 1 のような Precision-Recall 平面に 1 点プロットすることができる．閾値を動かすことにより，Precision-Recall の点をつなぐと曲線が得られる．その曲線は *Precision-Recall* **曲線**と呼ばれる．$(\lambda, m) = (0.1, 0.1)$ のときの Precision-Recall 曲線は，図 1 の実線のようになった．図 1 の点線と実線の交点は *Precision-Recall Break Even Point*(PRBEP) と呼ばれる．$(\lambda, m) = (0.1, 0.1)$ における PRBEP は $(0.9, 0.9)$ となった．(λ, m) の組み合わせを $\lambda = 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}$ および $m = 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}$ から網羅的に試したが，PRBEP が $(0.9, 0.9)$ より右上に移動することはなかった．

5 おわりに

本論文では，リアルタイムで入手できる水質水門データから大腸菌濃度をどれほど予測できるのか，2 クラス分類問題にして検証した．その結果，PRBEP が 0.9 となり，良好な予測性能が得られることが分かった．

謝辞：本研究は JSPS 科研費 26249075, 40401236 の助成を受けたものである．

参考文献

[1] T. Kato et al. Estimation of concentration ratio of indicator to pathogen-related gene in environmental water based on left-censored data. *Journal of Water and Health*, Vol. -, No. -, pp. xx–xx, - 2016. in press.