機械学習による水道水質内の大腸菌の存在予測-ネパールを事例に

－ Presence prediction of E. coli in water taps using machine learning in Nepal－

47-206753, 黒木颯
指導教員:坂本麻衣子

機械学習、水質、GIS、ネパール

1. はじめに

発展途上国では、水源から水道までの供給過程で、さまざまな問題が発生し、水道の汚染が生じる。安全な飲み水の確保を実現するためには、各地域の飲み水の汚染状況を把握する必要がある。しかし、WHO と UNICEF による Joint monitoring program は、世界規模で安全に管理された飲料水への家庭単位でのアクセス状況を推計して発表しているが、すべての国における安全に管理された飲料水にアクセスできている家庭を推定するために必要な、推計対象人口の 50%のデータを収集できていない(WHO/UNICEF, 2019)。本研究では、ネパール国内 26 都市で採取した 202 の水道水サンプルについて、サポートベクターマシン(SVM)、ランダムフォレスト(RF)、LightGBM の 3 つの機械学習モデルによる水道水中の大腸菌の有無の予測を検討する。発展途上国でよくみられることだが、入手可能な水質データや関連データは限られている。このような制約の中で、どのようにして正確な予測モデルを作成することができるかを検討し、また、作成した予測モデルをどのようにして現地の状況下で有効に活用することができるかを考察する。

2. 使用データ及び分析手法

本研究では、分析対象データとして、2013-2014 年度にネパール政府が日本国際協力機構(JICA)の支援を受け、他の都市のベンチマークとなるような大都市や歴史ある 26 都市を対象に実施した(Ministry of Health, 2016)、水源および水道水のサンプル採取と水消費者への聞き取り調査の結果を用いる。当該調査では、カトマンズの 21 の主要水源と無作為に選択された 103 の水道水、および他の 25 都市のすべての水源と無作為に選択された水道水を採取し、それぞれのサンプルについて pH、濁度、遊離残留塩素、大腸菌、電気伝導度などの水質が測定された。また各都市で無作為に選ばれた 50 世帯に聞き取り調査が行われた。聞き取り調査では年間の無水給水日の日数、1 日で水供給が行われる時間等が質問された。

2.1. 目的変数

本研究では、水の安全性を示す重要な水質指標である、大腸菌の存在有無を予測する。人々はその汚染を認識できないため、ろ過や煮沸などの適切な処理をせずにそのまま摂取してしまうという報告がなされている(Ogata et al., 2019)。ネパールの水道水質基準(NDWQS) において、大腸菌は検出されないことが基準となっている。したがって、本研究でも水質基準にのっとり、大腸菌を含まないことを基準にサンプルの分類をする。26 都市から取得した 202 の水道サンプルを大腸菌の検出有無で分類した結果、大腸菌を含まない水道が 92 個、大腸菌を含む水道が 110 個となった。

2.2. 説明変数

水源から水道への配水ネットワークを想定して説明変数を取得する。過学習を抑制するために水道水中の大腸菌の存在に特に関連性が高いと思われる 8 変数に絞り込んだ。8 つの変数は Type of the water source: 表流水か地下水、Number of processors: 都市別に実施される浄水処理方式の数、Amount of E. coli in the water source: 水源の大腸菌の数、Length: 水源または上流の水道からの距離、Days of no water: 年間の無給水日数、Number of buildings: 水源から水道への経路から作成した 100m 幅のバッファー上にある建物数、Population density: 経路周辺の人口密度、Agriculture area: バッファー上の農地面積、である。

2.3. 疑似水道ネットワーク

いくつかの説明変数のデータを得るためには、水源から水道までの水道管に沿った最短経路を得る必要がある。しかし、発展途上国では水道管の位置情報は容易に入手できないため、道路の位置情報を用いて、水道管が道路の下に埋設されていると仮定し、擬似的水道ネットワークを作成した。まず各道路の交差点を取得する。これを以下、道路ポイントと呼ぶ。次に各水源と水道を直線距離で最短な道路ポイントと紐づけ、各水源から水道までの最短経路を Dijkstra アルゴリズムを用いて求める。得られた最短経路をもとに、Type of the water source、Amount of E. coli in the water source、Length を得る。 さらに、最短経路から 100m 幅のバッファーを作成し、建物データを重ね合わせ、各バッファーと重なる Number of buildings をカウントする。同様に、各バッファーに土地利用図を重ねてバッファー上の Agriculture area を算出する。

2.4. 予測モデル

本研究では、予測モデルの手法として SVM、RF、LightGBM を用いた。SVM は 1995 年に発表された手法で、各クラスから超平面への距離を最大化することにより回帰・分類を行う。RF は 2001 年に発表された手法で、複数の決定木を組み合わせたものである。LightGBM は、2017 年に Microsoft が公開したアルゴリズムである。

3. 結果

3.1. 疑似水道ネットワーク分析の妥当性

推定した擬似水道ネットワークの妥当性を評価するために、取得した経路の上流と下流の大腸菌数を比較した。水道ネットワークが適切に取得されていれば、下流側の経路の大腸菌数が上流側の経路の大腸菌数より少なくなることはないと考えられる。各水源から各水道までの最短経路を Dijkstra で計算すると、上流側の大腸菌数を含む経路が 59 本得られた。カトマンズを除く 25 都市では、上流で大腸菌が検出された経路には必ず下流でも大腸菌が検出され、現実的な結果であると考えられる。一方カトマンズでは、上流で大腸菌が検出され、下流で大腸菌が検出されなかった経路が 26%存在した。

3.2. 予測モデルの作成及び精度評価

本研究では全 202 サンプルのうち 148 サンプルをトレーニングデータとし、残りの 54 サンプルをテストデータとして使用した。トレーニングデータに対して K-fold を 5 に設定しグリッドサーチを行うことでモデルの汎化性能が高まるようパラメーターの選択をした。テストデータでの予測結果は SVM、RF、LightGBM でそれぞれ 70%、61%、57%となり SVM の精度が最も優れていた。

3.3. 説明変数評価

SVM の説明変数を shap により評価した。結果から Type of the water source、Length、Days of no water が重要であることが分かった。水源が表流水であること、水源または上流の水道からの距離が長いこと、無給水日数がすくないことは、予測モデルが水道水内に大腸菌が存在すると予測する可能性を高めていた。

4. 結論

SVM が 70%の精度で水道水中の大腸菌の存在予測できることを確認した。また、SVM を用いた shap による説明変数の評価から、Type of the water source、Length、Days of no water が予測において重要な変数であることが分かった。疑似水道ネットワークの妥当性の分析では、カトマンズを除く 25 都市の結果は、上流と下流の関係を正確に捉え直感的な結果を示した一方で、カトマンズでは、上流で大腸菌を含むが下流では大腸菌を含まない経路が得られ、分析がうまくいっていない可能性を示した。本研究では得られた 70%の精度の予測モデルの活用例として、水道水を塩素で消毒する際に、大腸菌の混入が予測される地域から実施することで、より効率的に消毒を行うことができる可能性がある。

参考文献:

Health, W. (2005). *Ministry of Physical Planning and Works Singhadarbar kathmandu National Drinking Water Quality Standards ,*

Ministry of Health, L. and W. (2016). *Capacity Assessment and Benchmarking*. *72*.

Ogata, R., Khatri, N., & Sakamoto, M. (2019). Illuminating utility benchmarking data with analysis and consumer feedback

WHO/UNICEF. (2019). Progress on household drinking water, sanitation and hygiene 2000-2017.