4M-05

下水処理データの関係式発見に関する研究

笠川 舞夢 *　佐藤 利哉 †　高岡 旭 ‡　塩谷 浩之 §
室蘭工業大学 工学部情報電子工学系学科 *　室蘭工業大学 工学部情報電子工学系学科 †
室蘭工業大学 しくみ解明系領域 ‡　室蘭工業大学 しくみ解明系領域 §

1　はじめに

近年，下水処理施設の管理の民間委託が進みつつある．そのなかで，技術者減少による技術継承への懸念や施設管理体制の多様化に対応するための運用の効率化などが問題となっている [1].

これらの問題を解決するためには施設管理の自動化・効率化が必要である．自動化には既存の理論だけではなく設備にあったルールをデータから抽出する機構が必要であるため，数法則ニューラルネットワークを利用した運転管理データの関係式の発見についての研究を行った．これによりデータ間の関係を多項式で表すことができるようになる．

本研究ではデータ間の既知の関係式が数法則ニューラルネットワークによって再現できることを確認した．

2　数法則ニューラルネットワーク

数法則ニューラルネットワーク [2] とはデータの関係式を多項式で表すニューラルネットワークである．事例集合を $\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)\}$ とする．ただし $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \cdots, x_{tn})$ を n 次元の入力ベクトル，y_t を $\boldsymbol{x}_t$ に対する目標出力値，m を事例集合の個数とし，さらに $x_{ti} > 0$ を仮定する．数法則ニューラルネットワークの重みベクトル (パラメータ) を $\boldsymbol{\Phi} = (\boldsymbol{c}, \boldsymbol{w}_1, \cdots, \boldsymbol{w}_h)$，出力を $z(\boldsymbol{x}_t; \boldsymbol{\Phi})$ とするとき，$\boldsymbol{x}_t$ に対する数法則ニューラルネットワークで表される数式は以下のとおりである．

$$\begin{aligned} z(\boldsymbol{x}_t; \boldsymbol{\Phi}) &= c_0 + \sum_{i=1}^{h} c_i \prod_{j=1}^{n} x_{tj}^{w_{ij}} \\ &= c_0 + \sum_{i=1}^{h} c_i \exp\left(\sum_{j=1}^{n} w_{ij} \log x_{tj}\right) \end{aligned}$$

ただし $c_i, w_{ij} \in \mathbb{R}$，$\boldsymbol{c} = (c_0, \cdots, c_h)$，$\boldsymbol{w}_i = (w_{i1}, \cdots, w_{in})$，$h$ は定数項を除く多項式の項数とする．ここで $n = 2$，$h = 2$ の数法則ニューラルネットワークを図 1 に示す．

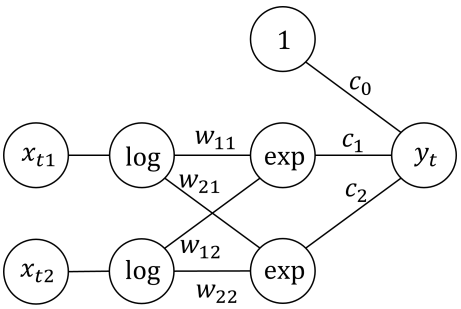

図 1: $n = 2$, $h = 2$ の数法則ニューラルネットワーク

損失関数を

$$f(\boldsymbol{\Phi}) = \frac{1}{2} \sum_{t=1}^{m} \{y_t - z(\boldsymbol{x}_t; \boldsymbol{\Phi})\}^2$$

とする．このとき $f(\boldsymbol{\Phi})$ の値が小さいほど実データとの誤差が小さいことを意味するため，良いモデルを得るためには $f(\boldsymbol{\Phi})$ を最小化する $\boldsymbol{\Phi}$ を求めればよい．そのための手法として BPQ アルゴリズム [3] を用いる．BPQ アルゴリズムは最急降下法と準ニュートン法を組み合わせたアルゴリズムである．それぞれを単体で使用したアルゴリズムよりも少ない反復回数で解を求めることができる．

3　モデル選択

数法則ニューラルネットワークの学習では項数 h をあらかじめ固定した状態で学習を行う．したがって各項数における最適なパラメータ $\boldsymbol{\Phi}$ を見つけ，さらにその中から最適な項数 h を見つけなければならない．

まずは各項数における最適なパラメータ $\boldsymbol{\Phi}$ を求める方法を説明する．BPQ アルゴリズムは解の収束がパラメータの初期値に依存するため，毎回パラメータの初期値をランダムに設定し，BPQ アルゴリズムを複数回行う．学習により得られた複数のパラメータの中から学習誤差が一番小さいものを各項数における最適なパラメータとする．ただし学習誤差は平均自乗誤差とする．

次に最適な項数を求める方法を説明する．項数が異なるモデルの場合，モデルの良さを学習誤差などでは比較できない．なぜなら項数が増えるとパラメータ $\boldsymbol{\Phi}$ の次元が増加し，それにより過学習を引き起こして未知データに対する推定の精度が悪

On finding relational expressions of sewage treatment data

* Maimu Kasakawa. Department of Information and Electronic Engineering, Faculty of Engineering, Muroran Institute of Technology.
† Toshiya Sato. Department of Information and Electronic Engineering, Faculty of Engineering, Muroran Institute of Technology.
‡ Asahi Takaoka. College of Information and Systems, Muroran Institute of Technology.
§ Hiroyuki Shioya. College of Information and Systems, Muroran Institute of Technology.

くなる可能性が高いからである．そこで先行研究ではMDL (Minimum Description Length) という評価尺度を用いてモデルの比較を行っている [2, 4]．本研究ではAIC (Akaike's Information Criterion) [5] を用いてモデルを比較する．AICの定義式は以下のとおりである．

$$\mathrm{AIC} = -2(\text{最大対数尤度}) + 2(\text{パラメータ } \boldsymbol{\Phi} \text{ の次元})$$

このときAICが一番小さくなるモデルが一番最適なモデルである．

ただし最大対数尤度はすべての学習データに対する誤差の同時確率密度関数の最大対数尤度とする．このとき誤差 $y_t - z(\boldsymbol{x}_t; \boldsymbol{\Phi})$ は平均0，分散 σ^2 の正規分布に従うと仮定する．

4 実験

今回実験に用いた深川浄化センターの運転管理データを表1に示す．今回の実験ではデータ間の既知の関係式を数法則ニューラルネットワークの学習により再現できるかどうか確かめるため，塩素注入率の関係式を他の7つのデータから再現できるか実験した．

今回の実験ではパラメータの初期値を決める際に平均0，分散1の正規分布を使用した．ただし c_0 の初期値は0に固定した．さらに数法則の各項数における最適なパラメータを求める際，BPQアルゴリズムは50回行った．学習には2015年の月報データを用いた．データは非負であり仮定を満たしている．

表 1: 深川浄化センターの運転管理データとその関係式

変数名	項目名	関係式
x_1	処理水量	-
x_2	塩素注入量	-
x_3	塩素注入率	$120x_1^{-1}x_2$
x_4	最初沈殿池沈殿時間	$10896x_1^{-1}$
x_5	曝気時間	$47280x_1^{-1}$
x_6	返送汚泥量	-
x_7	返送汚泥率	$100x_1^{-1}x_6$
x_8	最終沈殿池沈殿時間	$18288x_1^{-1}$

実験の結果得られた数式は以下のとおりである．

$$x_3 = 1.21x_1^{-0.0270}x_2^{1.00}x_4^{-0.181}x_5^{0.105}x_6^{-0.988}x_7^{0.988}x_8^{0.0610}$$

表1の関係式をもとに変形すると

$$x_3 \approx 120x_1^{-1.00}x_2^{1.00}$$

となり，塩素注入率の関係式を高い精度で再現できていることがわかる．

また未知データに対する推定精度を調べるために，2018年の月報データを用いて汎化誤差を調べた．ただし汎化誤差は平均自乗誤差とした．この時の汎化誤差は

$$\frac{1}{365}\sum_{t=1}^{365}\left\{y_t - z(\boldsymbol{x}_t; \hat{\boldsymbol{\Phi}})\right\}^2 \approx 2.48 \times 10^{-10}$$

となった．ただし $\hat{\boldsymbol{\Phi}}$ は数法則ニューラルネットワークの学習で最終的に得られたパラメータである．上の結果から汎化誤差も非常に小さいことがわかる．

5 まとめと今後の課題

今回はデータ間の既知の関係式の再現に成功した．今後は未知の関係式を発見することを目標とする．

しかし問題点がいくつか存在する．深川浄化センターの運転管理データは数百項目から構成されており，そのすべてを使って関係式を作るのはBPQアルゴリズムや他のアルゴリズムでも非現実的であるため，数法則ニューラルネットワークに用いる変数の数を減らさなければならない．そのため，使用する変数は関係式を求めたいデータと相関があるデータのみにしたい．そこで，データ間の相関をどのように見つけるかが次の問題となる．今候補にあるのは相互情報量を使うということだが，多変数で相関があるようなデータの場合，相互情報量だけでは相関を見つけられないデータの組み合わせが存在する可能性がある．今後はどのように変数の数を削減するか，どのようにデータの相関を見つけるかが問題となる．

本研究は，北海道における数理データサイエンスによる地域貢献につながる研究として位置付けている．研究実施において，実際の下水処理施設に関するデータ事項や助言などを頂いた株式会社データベースに深謝する．

参考文献

[1] 国土交通省水管理・国土保全局下水道部．下水道事業の現状と課題―持続可能な下水道事業とするために―．2019. https://www.mlit.go.jp/mizukokudo/sewerage/content/001313228.pdf

[2] 斉藤 和巳，中野 良平．コネクショニストアプローチによる数法則の発見．情報処理学会論文誌．vol.37，no.9，pp.1708–1716，1996.

[3] K. Saito and R. Nakano. Partial BFGS update and efficient step-length calculation for three-layer neural networks. Neural computation，vol.9，no.1，pp.239–257，1997.

[4] 大澤 知弥，塚本 蔵人，高岡 旭，塩谷 浩之，柳本 光皓，日置 岳彦，大森 康弘．数法則ニューラルネットワークによる下水処理施設の運転管理効率化の検討．情報処理学会北海道シンポジュウム2020予稿集，頁187-188，2020.

[5] 坂元 慶行，石黒 真木夫，北川 源四郎．情報量統計学．情報科学講座 (A・5・4)，1983.