(57) 重回帰分析とディープラーニングの比較

吉永　弘志[1]

[1]正会員　国立研究開発法人　土木研究所　主任研究員（〒305-8516 茨城県つくば市南原 1-6）
E-mail: h-yoshinaga@pwri.go.jp

重回帰分析等の統計的手法よりも Deep learning によるデータ解析が優れている旨の報告事例が散見されるようになった. そこで, 騒音, 振動, および粉塵のデータ解析への Deep learning の活用を目的としてダミーデータ, および測定データを分析した. Deep learning は，内挿は優れているが 外挿の推計は直線的であること, 異常データの影響を受けやすいこと, および説明変数を三角関数等で変換して目的変数との関係を線形にできる場合には重回帰分析に劣るが線形にできない場合には優れていることを把握した. 既存のデータを再解析した二つの事例では, 既報の方法に対する優位性を見出せなかったが, 少ない数の測定値から目的変数を最大にする説明変数を見出すこと, および発生源を推定することへの応用の可能性を見出した.

Key Words: *Deep learning, multiple linear regression, prediction, noise, dust, DFT*

1. はじめに

2012 年の画像認識のコンテスト ILSVRC において他を圧倒して優勝したことを契機にディープラーニング (以下,「Deep learning」という.) の活用が拡大している. 土木事業の騒音, 振動, および粉塵の環境影響評価では現場での測定値を回帰分析等の統計的な手法で解析して整理したデータを使用しているが, 将来は Deep learning 等の AI でデータ解析することも考えられる. 建設工事による PM10 の予測に AI を活用した論文[1]がみうけられるようになり, 建設工事の粉塵予測に活用する検討をしている企業もある. 近年は AI の性能の向上と普及が著しく, 医療のヘルスケアのデータ解析[2], およびビルのエネルギー消費の予測[3]においては重回帰分析よりも Deep learning の方が優れていた旨の報告例もみうけられるようになった.

しかし, Deep learning 等の AI での計算はブラックボックスであること, 日進月歩でアルゴリズムが進化していること, 様々なソフトウエアが乱立していること, および学習 (教師) データやアルゴリスムの設定等の使い方で性能に大きな違いが生じることから, 特性をよく把握した適材適所の活用が重要であると考えている.

そこで, 簡易な関数で測定値を模擬したデータ (以下,「ダミーデータ」という.) で Deep learning の特性を把握し, 過年度の現場での測定値を重回帰分析と Deep learning で再解析した.

2. 方法

(1) ソフトウエア

Deep learning には米国の H_2O.ai 社が無償で公開していた H_2O を使用した. ビジネスマン向けの講習会で紹介されており, インターネットブラウザのクリック操作で AI の解析ができる. H_2O はアルゴリズムを複数試行し, 中間層の層の数と各層でのノード数を自動で最適化する. 他のアルゴリズムも選択できるが, 広く一般に知られるようになった Deep learning のみ使用した

なお, 重回帰分析では統計解析ソフト R を使用し, 離散フーリエ変換 DFT はマイクロソフトのエクセルに計算式を入力して計算した.

⑵ データの分類

データは**図-1** のように分類した. 目的変数の値のある学習用データと目的変数を空欄にした評価用データに分け, Deep learning では学習用データを tarainig dataset, および

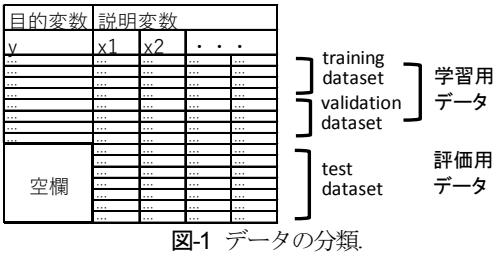

図-1 データの分類.

validation dataset に分けた. 評価用データでは推計値（本稿では計算で予測した値を「推計値」と称した.）と正解（想定した関数, および測定値）を比較した.

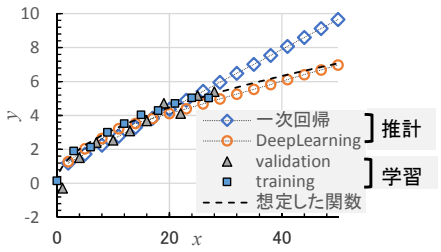

図-2 $y = x^{\frac{1}{2}}$を模擬したデータの解析.

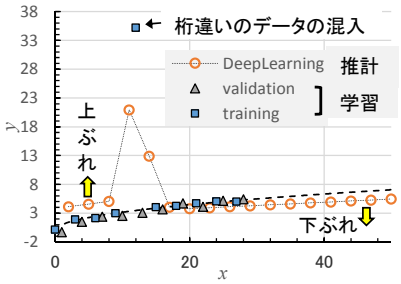

図-3 桁違いのデータの混入を想定した推計.

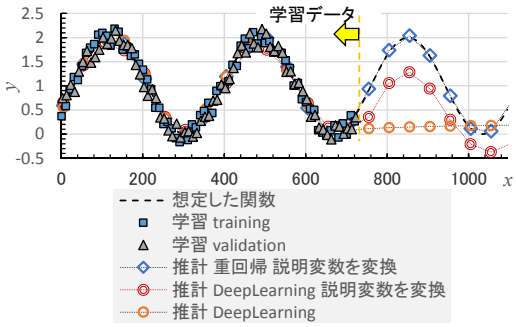

図-4 $y = \sin(2\pi(x-30)/365) + 1$の模擬データの解析.

(1) 設定した関数　　(2) 学習用データ

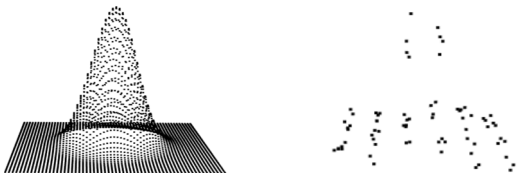

図-5 $z = e^{-(x-3)^2}e^{-(y-3)^2}$を模擬したデータ.

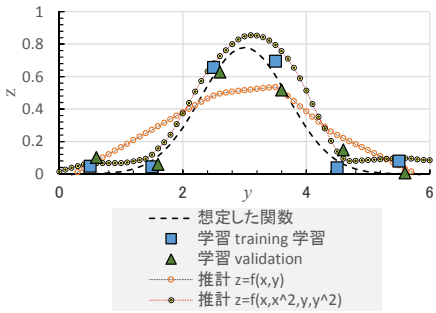

図-6 $z = e^{-(x-3)^2}e^{-(y-3)^2}$を模擬したデータの解析.

(3) ダミーデータの解析

数式でデータを設定した. データには乱数で模擬的に偶然誤差を付加した.

(4) 測定値の解析

過年度に報告した騒音 [4], および粉塵 [5]の測定値を再解析した. 重回帰分析および Deep learning での推計結果は測定値Mと推計値Pの差eの平均値$\bar{e}$(式(1)) で評価した.

$$\bar{e} = \overline{|M - P|} \qquad (1)$$

3. データ解析

(1) ダミーデータ

a) $y = x^{\frac{1}{2}}$

関数形に関する情報がない前提で説明変数xと目的変数yの関係を解析した結果を図-2 に示す. 一次回帰では$30 < x$の外挿での乖離が大きくなる. Deep learning では$x < 30$で曲線をよく模擬しており$30 < x$の外挿でもおきな乖離が生じていない. 次に tarainig dataset に桁違いのデータが混入した場合を想定し Deep learning で推計した (図-3). 内挿, 外挿ともに推計値が乖離する.

b) $y = \sin(2\pi(x-30)/365) + 1$

365 日周期で変化するデータを想定した解析結果を図-4 に示す. Deep learning は説明変数$x < 720$の内挿では曲線をよく模擬しているが$720 < x$の外挿では直線的な推計になり, 大きく乖離する. そこで周期 365 日は既知を前提とし, 説明変数xを変換した$x1 = \sin(2\pi x/365)$と$x2 = \cos(2\pi x/365)$を説明変数に追加した. 重回帰分析での推計はほぼ想定した関数と一致し, Deep learning の推計も改善された.

なお, 周期 365 日が未知の場合を想定して学習データから離散フーリエ変換 DFT で周期を分析したら 373 日周期となった. 付録で補足説明する.

c) $z = e^{-(x-3)^2}e^{-(y-3)^2}$

説明変数(x, y) = (3, 3) で目的変数zが最大になるガウス分布を模擬した (図-5). 学習用データは図-5 の(2)のように設定した. ガウス分布が既知の場合には,

$$\ln z = ax^2 + bx + cy^2 + dy + e \qquad (2)$$

と変換して係数 $a, b, c,$ および d を重回帰分析する方法が簡便で正確であることが b)の事例よりわかるが, ここでは関数形が未知の場合を想定して Deep learning のみで解析した. $x = 3.5$での解析結果を図-6 に示す. 説明変数を変換しない$z = f(x, y)$として解析した場合はいびつな推計になるが, 説明変数にx^2, およびy^2追加すると推計

の性能が向上しガウス分布を想起することができる. また, 推計ではzが最大になる(x, y)は(3.1, 3.1)となり設定した関数とほぼ一致した.

(2) 測定値

a) 騒音

公道を走行する大型車420台から発生する騒音L_{WA}, 速度V, および質量Mの測定値を重回帰分析し, dB 単位のL_{WA}を$P_A = 10^{LWA/10} \cdot 10^{-9}$で mW に換算した値$P_A$が$V$の 3 乗, および$M$の 0.5 乗に比例していた旨の報告[4]をしているが, ここでは測定値を学習用データ210台と評価用データ 210 台に分けて重回帰分析, および Deep learning で再解析した. Deep learning での tarainig dataset, および validation dataset の数は双方とも 105 台にした. 推計値を式(1)で評価した結果を表-1 に示す. 測定値のままでの解析では重回帰分析より Deep learning の方が優れていたが, 説明変数V, およびMを対数で変換した$\ln V$, および$\ln M$を説明変数とした解析では重回帰分析の方が優れていた. この重回帰分析の方法は既報でも採択していた.

表-1 騒音の解析結果の比較.

	誤差の平均 dB	
重回帰 測定値のまま解析	1.25	
重回帰 速度Vおよび質量Mを対数にして解析	1.14	◎最小
Deeplearnig 測定値のまま解析	1.21	
Deeplearnig 速度Vおよび質量Mを対数にして解析	1.32	

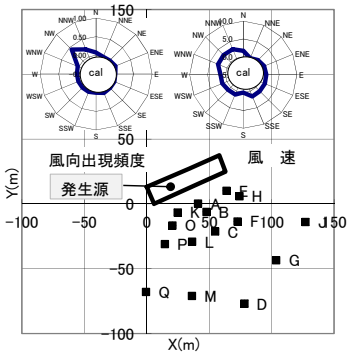

図-7 粉塵の測定. A～Q は測定点.

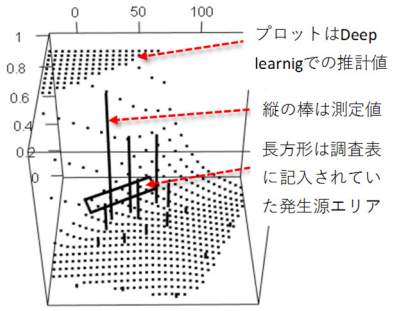

図-8 測定値に基づく Deep learnig での推計値. 上限値を超える推計値は上限値でグラフ化.

表-2 降下ばいじんの再解析結果.

	誤差の平均値	備考
既報の解析	0.073	アセス用データに反映済
Deep learning	0.062	他で流用できない
既報の方法で発生源位置修正	0.062	

b) 粉塵

土木工事に起因して発生する粉塵を降下ばいじんとして測定した結果を整理して環境影響評価の予測用データとして報告している[5]が, ここでは粉塵の発生量が多い路床安定処理のなかから一つの現場 (図-7) を選定して再解析した. 既報では

$$C_{\mathrm{d-gound}} = \sum_{\mathrm{directions}} \iint_R a A^{-1} u^{-1} x^{-2} \mathrm{d}S \qquad (3)$$

を仮定し, 降下ばいじんの測定値$C_{\mathrm{d-gound}}$ (t / km^2・日), 発生源エリアの面積A (m^2), 風向 directions, 風速u (m / s), および発生源領域 R のなかの微小領域dS から測定点（予測地点）までの距離x (m) から定数aを悉皆的に調べる方法で算出していた.

式(3)を前提として重回帰分析, および Deep learning を適用する方策が不明だったため, 説明変数を発生源の中心からの距離L, および最頻風向での風下方向との角度θとの余弦$\cos\theta$として Deep learning で推計したら既報よりも誤差が小さくなった. しかし, 風向がずれると降下ばいじん量が増えるとの現実にはありえない推計値になった.

そこで他目的での活用を想定した解析を行った. 既報[5]では発生源エリアを修正すると推計誤差が小さくなる現場がいくつかあった. そこで, 発生源エリアの確認に活用することを試行した. 図-8 は測定値, Deep learning での推計値, および調査表に記入されいた発生源エリアである. 逆二乗で距離減衰する式(3)よりも急な減衰をしているようにみうけられたので, 平行移動して再解析したら 誤差の平均が小さくなった(表-2).

なお, 説明変数を座標とした Deep learning での推計でも誤差は小さくなるが(表-2), 他の現場には活用できない. また, 距離xのべき数も説明変数として解析すると$x^{-3.2}$となり誤差は小さくなるが, 複数の現場を俯瞰してx^{-2}で統一していた[5].

4. 考察

Deep learning には以下の特性があると考えた.

・内挿的な推計の性能は高いが, 学習データに誤りのデータが混入した場合は影響を受けやすい.

・説明変数 (変換した説明変数も含む) と目的変数が線形（直線）の関係の場合は Deep learning よりも重回帰分析の方が優位であり, 非線形（曲線）の場合は, Deep learning が優位である.

さらに Deep learning は少ない数の測定値に基づく以下に応用できると考えた.

・推計に整合する関数の推定，およびコスト最小，効率最大等の最適なパラメータを予測すること(図-6).

・発生源の位置推定(図-8).

最後に以下に留意することが必要と考えた. Deep learning では過学習とされている現象を避けるため, 学習の過程で validation を行う対策が講じられているが, 万全とは言い難い. 高い能力を過信することなく人の判断で賢明な使い方をすることが重要である.

5. まとめ

騒音, 振動, および粉塵等のデータ解析に活用することを目的として重回帰分析と Deep learning を以下で比較した.

・ソフトウエアは H_2O.ai 社の H_2O.

・データは各種の関数で設定したダミーデータ, および現場での測定値.

Deep learning には以下の特性があると考えた.

・説明変数と目的変数に曲線の関係がある場合でも内挿は優れているが, 外挿での推計は学習データの直線的な延長 (図-4).

・特異的なデータの影響を受けやすい (図-3).

・正弦波やガウス分布のように増減するデータは測定値のままでは解析できないので説明変数を三角関数, 二次関数等で変換することが必要（ 図-4, 図-6 ）.

・説明変数 (変換した説明変数も含む) と目的変数が線形（直線）の関係の場合は Deep learning よりも重回帰分析の方が優位であり，非線形（曲線）の場合は，Deep learning が優位.

Deep learning で既存の騒音, および粉塵のデータを再解析した結果, 以下となった.

・騒音のデータでは重回帰分析の方が優位(表-1).

・粉塵のデータでは誤差を小さくする推計もできたが実務に活用する見込みはなかった.

Deep learning は少ない数の測定値に基づく以下の活用ができると考えた.

・関数形の推測, および目的変数が最大・最小になる説明変数の推定(図-6).

・発生源の推定(図-8)

最後に, Deep learning は, 学習用 (教師) データに不良なデータが混入しないように留意すること等, 賢明な使い方が重要であること, および着目されることになった契機が性能のコンテストであったことは, 人間社会にも通じる点があると感じたことを付記する.

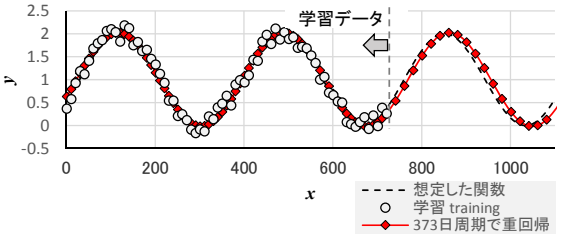

図-9 離散フーリエ変換 DFT での周期分析と重回帰分析の組み合わせによる周期的なデータの分析例.

付録　周期の分析

図-4 のデータ解析では周期 365 日を前提としたが周期が未知の場合を想定して離散フーリエ分散 DFT で周期Tを分析した. 周期Tの出力$F(T)$をn番目のデータ, およびデータの平均値を$f(n)$, および$\bar{f}$として,

$$F(T)=\left|\sum_{n=1}^{N}(f(n)-\bar{f})e^{-i\frac{2\pi n}{TN}}\right| \qquad (4)$$

で計算したら,$F(T)$が最大になる周期Tは 373 日になった. 図-9 に示すように周期 373 日とした重回帰分析での推計値は想定した関数とほぼ一致する.

なお, 周波数分析では高速フーリエ変換 FFT を使用するのが一般的であるが, 使用するデータを 10 日間隔で 64 個 (FFT は 2 の累乗数のデータを使用するため) としたら分析できる周期は 213, 320, 640 等の飛び飛びの値となり, 出力の最大として推計した周期が 320 日となったので DFT で分析した.

参考文献

1) Feng Q, Wu SJ, Du Y, Xue HP, Xiao F, Ban X, Li XD: Improving Neural Network Prediction Accuracy for PM10 Individual Air Quality Index Pollution Levels, *ENVIRONMENTAL ENGINEERING SCIENCE*, vol.30, 12, pp.725-732, 2013.
2) Gurupur VP, Kulkarni SA, Liu X, Desai U, Nasir A: Analysing the power of Deep learning techniques over the traditional methods using medicare utilisation and provider data, *JOURNAL OF EXPERIMENTAL & THEORETICAL ARTIFICIAL INTELLIGENCE*, vol.31, No.1, online, 2019.
3) Li CD; Ding ZX; Zhao DB; Yi JQ; Zhang GQ: Building Energy Consumption Prediction: An Extreme Deep learning Approach, *ENERGIES*, vol.10, No.10, online, 2017.
4) 吉永弘志，大河内恵子, 井上隆司：公道を定常走行する大型車の質量・速度・騒音の測定値, 自動車技術会論文集, Vol.46, No.4, pp.781-786, 2015.
5) 山元弘, 林輝, 吉永弘志, 吉田潔: 建設工事騒音・振動・大気質の予測に関する研究(第 3 報), 土木研究所資料第 4010 号, 2006.