直線データに対する正規圧縮距離へのガウス型白色ノイズの影響

石原正道†

†郡山女子大学人間生活学科

1　はじめに

今日、利用できる情報は極めて多く，また日毎に情報量は増加している．このため現在では情報をいかに取得するかという点だけでなく，いかに活用するかという点も重要な課題となっている．情報を活用するための手法であるデータマイニングでは，情報の適切なグループ化・ラベル付け・要約などが行われる．グループ化の一手法に距離行列法があり，なんらかの意味で定義した距離を用いてグループ化を行う．このため距離行列法においては距離をどう定義するか，また定義した距離からどのようにグループ化していくのかが問題となる．

この距離として，近年コルモゴロフ複雑量を基礎とする正規圧縮距離が提案された．正規圧縮距離においてはコルモゴロフ複雑量を圧縮した後のデータ量で近似する．これまで正規圧縮距離は文学 [1, 2] や音楽 [1, 3] など様々な種類のデータの分類に適用され，この距離が分類に有効であることが示されている．文学を記述する際に用いられる文字，音楽データの記述形式であるMIDI，またゲノムデータなどは程度の差異はあるものの離散データといってよい．離散的なデータはノイズが作用してもノイズの影響を除去しやすい．このため離散データはノイズに影響されにくいと考えられる．したがって離散データに対する正規圧縮距離もノイズに影響されにくいと考えられる．一方で音声や映像は本来連続的なデータであり，これらのデータにもノイズは作用しうる．このため連続データにノイズが作用した際に正規圧縮距離がどのような影響を受けるのか知っておく必要がある．

これまでの研究において，擬ランダムなノイズが印加されたデータを圧縮した際のデータサイズは，非常に大きくなるということが指摘されている [4]．一方で，ノイズが印加されているデータにおいて正規圧縮距離によるグループ化が有効に働きうることが指摘されている [5]．またノイズが印加された連続データに対する正規圧縮距離の挙動も調べられている [6]．

様々な現象を記述する際の基本的な関数には指数関数および三角関数が用いられる，分析などで関数を近似する際には一次関数がよく用いられる．そこで本文では，データが一次関数により生成される場合において，正規圧縮距離はノイズにどのように影響を強く受けるのかということについて調べた結果を報告をする．

2　正規圧縮距離とノイズを含むデータ

まず正規圧縮距離を定義する．ファイルを P，Q とし，$C(P)$，$C(Q)$ を圧縮した後のファイルサイズとする．このとき正規圧縮距離 (NCD) は次式で定義される [7, 8]:

$$\mathrm{NCD}(P,Q) := \frac{\max\left[C(PQ)-C(P), C(QP)-C(Q)\right]}{\max\left[C(P), C(Q)\right]}. \tag{1}$$

ここで，$\max[x,y]$ は x,y のうち大きい値をとる関数である．次にデータの作成方法について述べる．パラメータ t に対し，ある値を返す関数 $x(t)$ を考える．t に関して区間 $[0,T]$ を考え，Δt 毎に分割し，各区間の先頭の値を t_i とする．この t_i を用いて，データを

$$x_j(t_i) = x(t_i) + Ag_j(t_i) \tag{2}$$

と構成する．ここで $g(t)$ は平均0，分散1の白色ノイズであり，j は $[0,T]$ における一つのノイズの列を指定している．また A は定数であり，数値 $x_j(t_i)$ は10進法で表わされているものとする．数値 $x_j(t_i)$ において，小数点以下第 $(k+1)$ 位を切捨てたデータを $x_j(t_i;k,A)$ と記す．ここではデータが A にも依存することも明示している．得られた $x_j(t_i;k,A)$ からなる列を蓄えたファイルを $X_j(k,A)$ とする．ファイル $X_j(k,A)$ とノイズを印加していないデータ $X(k,0)$ との間の正規圧縮距離を $\mathrm{NCD}(X_j(k,A),X(k,0))$ と記す．ノイズが印加されているために得られた距離にばらつきがでるから、ノイズの系列に関する算術平均をとる。より具体的には添字 j について算術平均をとり、この量を $\langle \mathrm{NCD} \rangle_{(k,A)}$

Effects of Gaussian White Noise on Normalized Compression Distance between Linear Data

Masamichi ISHIHARA†

†Dept. of Human Life Studies, Koriyama Women's University 963-8503, Kaisei 3-25-2, Koriyama, Japan

m_isihar@koriyama-kgc.ac.jp

と記す．ノイズの影響の度合をみる方法の一つは A を固定して $\langle \mathrm{NCD} \rangle_{(k,A)}$ の k 依存性をみることである．本研究では $\langle \mathrm{NCD} \rangle_{(k,A)}$ に対するノイズの影響を k 依存性により調べた．

具体的にデータを構成するためには関数 $x(t)$ や変数 t を分割する区間数を定める必要がある．またこれらのデータから $\langle \mathrm{NCD} \rangle_{(k,A)}$ を求めるには圧縮ソフトを定める必要がある．そこで本計算では区間数は 500 とし，圧縮ソフトとして bzip2 を用いた．区間は $[0,1]$ をとっている．従ってステップ幅 Δt は 0.002 となる．また本論の目的から、$x(t)$ としては切片 0 の一次関数をとることとする。すなわち

$$x(t) = a_s t \tag{3}$$

ととることとした。ここで添字 s により数値 a_s は次のように決められる。ある数 a を考える。この数値の小数点以下第 s 桁までの数値を a_s とする。a_s の小数点以下第 $(s+1)$ 位からは 0 が設定されるものとする。s を変動させることで、係数 a_s に対する応答をみることができる。

3 計算結果

本節では式 (3) により生成したデータを用い、計算した正規圧縮距離の結果を示す。

まず小数点以下の桁数 k と NCD の平均値は $\langle \mathrm{NCD} \rangle_{(k,A)}$ の関係について得た結果を述べる。これまでの計算結果 [6] から $x(t) = \sin(t)$, $\exp(-t)$ などの場合においては、一般的に k が増加するにしたがって $\langle \mathrm{NCD} \rangle_{(k,A)}$ の値は増加する。とくにノイズの影響を極わずかでも含むようになると、$\langle \mathrm{NCD} \rangle_{(k,A)}$ の値は急激に増加する。一方で一次関数 $x(t) = a_s t$ において a をネイピア数にとった場合[1]、$\langle \mathrm{NCD} \rangle_{(k,A)}$ はある k でピークを持ち、その後は k の増加とともに減少していく。またピークの位置での k の値は s 比例していることも分かる。ピークの位置は s の値 $(s \in \mathbf{N})$ と時間分割幅 Δt によって概ね定まるようである。同様の傾向は a として他の値 (例えば円周率 π) をとった場合にも見られる。

4 まとめ

本論では一次関数により生成され，かつ，このデータにガウス型白色ノイズが印加されたデータに着目した．得られたデータに対して正規圧縮距離を計算し，正規圧縮距離がどのような振る舞うかについて調べた．

通常はノイズの作用が顕著になり始めると正規圧縮距離は著しく増加する．これに対し直線データの場合はノイズの影響を受けやすくなるにも係わらず正規圧縮距離の値が減少をすることがわかった．このことは正規圧縮距離を数値データ適用する際に，安易な線形化を行うと正しい距離を算出できない可能性があることを示唆するものである．本研究は数値データへの正規圧縮距離を適用方法を理解するための一助となるだろう．

[1] $s = 0$ に対しては $a_0 = 2$(実数) ととるものとする。

参考文献

[1] Cilibrasi, R. and Vitányi, P.: Similarity of Objects and the Meaning of Words. arXiv:cs/0602065.

[2] 石原 正道 佐藤 静香: 正規圧縮距離を用いた和文小説の著者別分類と圧縮プログラムの妥当性, 情報処理学会論文誌 Vol. 49, No. 12, pp. 4016–4024 (2008).

[3] Cilibrasi, R., Vitányi, P. and de Wolf, R.: Algorithmic Clustering of Music. arXiv:cs/0303025v1.

[4] D. Sculley and Carla E. Brodley: Compression and Machine Learning: A New Perspective on Feature Space Vectors, *DCC*, pp.332-341 (2006)

[5] M. Cebrián, M. Alfonseca and A. Ortega: The Normalized Compression Distance Is Resistant to Noise, *IEEE Trans. Inf. Theor.* Vol. 53, No. 5, pp.1895-1900 (2007)

[6] 石原 正道: 正規圧縮距離に対するガウス型白色ノイズの影響，第 71 回情報処理学会全国大会講演論文集, 1-249-1-250 (2009).

[7] 渡辺 治: 計算機から見たランダムネス，統計数理, Vol. 54, No. 2, pp. 511–523 (2006).

[8] M. Cebrián, M. Alfonseca and A. Ortega, "Common pitfalls using the normalized compression distance: What to watch out for in a compressor," *Communications in information and systems*, vol. 5, no. 4, pp. 367–384 (2005).