正規圧縮距離に対する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

利用できる情報量は日々増加しているが，これらの情報を活用するには適切なグループ化・ラベル付け・要約などが必要となる．グループ化の方法の一つに距離行列法があり，グループ化をする際の強力な手法となっている．距離行列法を利用するにあたっ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かが問題となる．近年，現実的な距離として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．この距離ではコルモゴロフ複雑量を圧縮したデータ量で近似するという方法をとっている．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有効であることが示されている．この分類に用いられた多くのデータは離散データである．離散データはノイズに影響されにくいと考えられるため，必然的にこれらの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いえる．一方でデータには連続データもあり，これらのデータにもノイズは作用しうる．正規圧縮距離に対するノイズの影響について，いくつかの点が指摘されている．具体的には，擬ランダムなノイズの影響を受けたデータを圧縮した際のデータサイズは，コルモゴロフ複雑量より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，ノイズがある際にも正規圧縮距離によるグループ化が有効に働くとも指摘されている [5]．従って，現時点においては連続データに対する正規圧縮距離がノイズの影響を受けるかという点について，明瞭になっているとは言い難い．そこで本文では，連続データに対し，正規圧縮距離はノイズの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．

Effects of Gaussian White Noise on Normalized Compression Distance
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

2 ノイズを含むデータに対する正規圧縮距離

まず正規圧縮距離を定義しておくことにする．データを蓄えたファイルを 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 [6, 7].

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

ここで，$\max[x,y]$ は x,y のうち大きい値をとる関数である．さてパラメータ t に対し，ある値を返す関数 $x(t)$ を考える．t の区間 $[0,T]$ を 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 はこのノイズによる一つの列を指定している．また A は定数である．さて $x_j(t_i)$ は10進法で表わされているものとする．この $x_j(t_i)$ 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 にも依存することを明示した．この $x_j(t_i;k,A)$ 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 $X_j(k,A)$ とノイズのない場合のデータ $X(k,0)$ との間の正規圧縮距離 $\mathrm{NCD}(X_j(k,A),X(k,0))$ とし，添字 j について算術平均を $\langle\mathrm{NCD}\rangle_{(k,A)}$ と記す．$\langle\mathrm{NCD}\rangle_{(k,A)}$ の定義から，A が大きいほど，また k が大きいほどノイズの影響を受けることが予想される．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．

上記の説明から分かるように，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．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．また区間は $[0,2\pi]$ をとっている．従ってステップ幅 Δt は $2\pi/500$ となる．

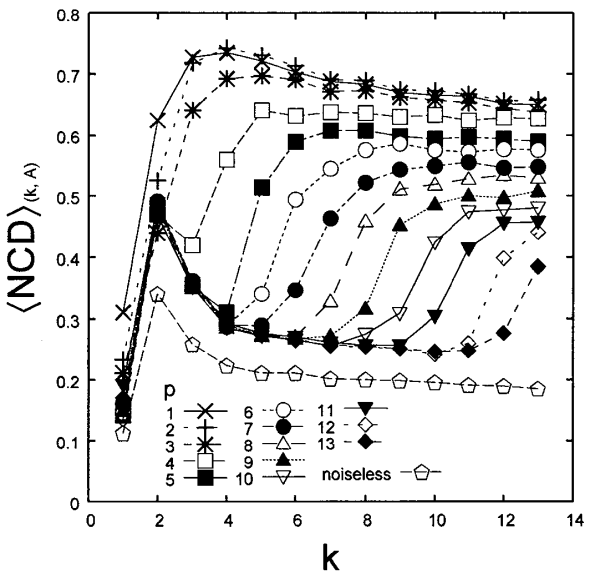

図 1: $x(t) = \sin(t)$ の場合における，小数点以下の桁数 k と NCD の平均値．係数 A は式 $A = 10^{-p}$ により p で表されている．

3 計算結果

本節では数値計算結果を示す．まず関数 $x(t)$ として正弦関数 $\sin(t)$ をとった結果を示す．図 1 は $x(t) = \sin(t)$ の場合であり，横軸は小数点以下の桁数 k を表し，縦軸は NCD の平均値を表している．また係数 A を $A = 10^{-p}$ と表し，p の値で係数の値を区別している．NCD の変動をみると p の値と同じ k の値のところで NCD は急速に大きくなっている．これはノイズの影響が効き始める桁数に対応しており，NCD はわずかなノイズの影響を強く受けていることを意味している．図 1 では $p = 1, 2$ において NCD の値が減少しているが，これはノイズのない場合の NCD の変動を反映している可能性がある．$x(t)$ を余弦関数や指数関数にとった場合も NCD は同様の k 依存性を有している．

同様の計算を一次関数に対して実行することができる．そこで $x(t)$ の一例として，$x(t) = 2.0t$ として計算を実行し，桁数 k と NCD の平均値の結果を得た．その結果は $x(t) = \sin(t)$ の場合と異なり，$\mathrm{NCD}(X(k,0), X(k,0))$ が k の関数としてほぼ一定である領域においても，$\langle \mathrm{NCD} \rangle_{(k,A)}$ には k の関数として明らかな減少傾向がみられた．

4 まとめ

本論では実数により表現されるデータに着目し，データに印加されたガウス型白色ノイズによって正規圧縮距離がどのような影響されるのかという点について調べた．また離散データとの対比も考慮し，実数の小数点以下の桁数を限定することで離散化とノイズの関係を調べた．

数値計算から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した．この結果から，正規圧縮距離はノイズの影響を受けやすいことが分かる．一方でノイズの印加されていないデータが (本文での) 一次関数である場合においては，正規圧縮距離はノイズにより値が大きくなるとは限らないこと分かった．この事から元のデータの種類によってはノイズが作用しても正規圧縮距離を増大させない場合のあることが分かる．ノイズによる正規圧縮距離の変化を詳細に調べることで，正規圧縮距離を用いた距離行列法の適用可能範囲の一端を知ることができるであろう．

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] 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).

[7] 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).