3B-7

正規圧縮距離によるノイズを含むデータの分類の可能性

石原正道 †

† 郡山女子大学人間生活学科

1 はじめに

利用できる情報が増大した現在，得られたデータをいかに整理し分類していくかということは重要な課題となっている．

データを活用をするための手法であるデータマイニングでは，何らかの意味での (非) 類似性を用いてデータの整理や分類を行う．したがって類似度あるいは距離を定義する必要があるが，その一つに正規 (化) 圧縮距離 (NCD) がある．

コルモゴロフ複雑量を基礎とする NCD では，データの情報量を圧縮後のファイルサイズで代用し距離を定めている．この距離を用いた分類例に文学 [1, 2] や音楽 [1, 3] などがあり，これまで種々のデータの分類に対して有効であることが分かっている．文学・MIDI データ・ゲノムデータなどは有限の文字種あるいは記号から構成されているから，これらのデータは離散データである．一方で音楽データや画像データは本質的には連続データであり，NCD を適用できるか自明でない．

離散データと連続データを比較した場合，離散データはノイズを除去しやすく連続データではノイズを除去しにくいといった特徴がある．このため連続データに対して NCD を適用する際に，ノイズが及ぼす影響が重要となりうる．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では圧縮後のデータサイズは大きいことが指摘されている [4]．さらに NCD はノイズの影響を強く受けるであろうことも示されている [5, 6].

NCD がノイズの影響を受けるならば，ノイズを有するデータを NCD で分類することは困難に思われる．しかしノイズの影響により，分類をされるデータが同じように圧縮しにくくなるのならば，データを分類できる可能性がある．またデータから適当な量を抽出することにより，データを分類できる可能性もある．実際，ノイズが印加されたデータにおいても NCD によってデータを分類できるとの報告もある [7].

以上の点を明確にするため，本研究では，時系列データにみられるような一次元的に記述されたデータは NCD を用いて分類が可能であるか調べた．実験などで得られるデータはノイズを有することが多いと考えられるため，自然科学で現れる典型的な関数から得られる数値に対しノイズを印加してデータを生成した．これらのデータ間の NCD を求め，多次元尺度構成法 (MDS) による分類やクラスタ分析による分類を試みた．また生成したデータから自己相関関数などの量を抽出し，得られたデータ間の NCD を求め，同様の分類を試みた．

2 正規圧縮距離とノイズを含むデータ

NCD を定義しておく．ファイルを P，Q で表し，$C(P)$，$C(Q)$ を圧縮した後のファイルサイズとする．また $\max[x, y]$ は x, y のうち大きい値をとる関数とする．これらの量と関数を用いて NCD は次式で定義される [8, 9]:

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

NCD の計算では具体的に圧縮ソフトを定める必要がある．本計算では圧縮ソフトとして bzip2 を用いた．

ファイル P，Q に含まれるデータは次のように構成した．まず変数 t に対し，ある値を返す関数 $x(t)$ を考える．変数 t の区間 $[0, T]$ を定め，この区間を Δt 毎に分割する．分割された各区間の先頭の値を t_i とするとき，データを

$$x_j(t_i) = x(t_i) + Ag_j(t_i)$$

により構成する．ここで $g(t)$ は平均 0，分散 1 の白色ノイズであり，j はあるノイズの列を指定している．定数 A はデータに取り込まれるノイズの強さを決めている．本研究においては数値 $x_j(t_i)$ は 10 進法で表わされているものとする．得られた $x_j(t_i)$ をファイルに保存し使用した．データを生成する関数 $x(t)$ は様々な関数が

The possibility of Classification of Noisy Data by 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

考えられる．ここでは自然科学で頻出する関数 $\sin(t)$, $\cos(t)$, $\exp(-t)$, t, t^2 をとった．また区間を定める T は 2π とし，分割幅 Δt を 0.001 とした．一つの関数 $x(t)$ につきノイズを含むデータを有するファイルを 10 ファイル，計 50 ファイル作成した．また $x_j(t_i)$ そのものをデータとするだけでなく，標本自己共分散関数および標本自己相関関数を求め，これらのデータを有するファイルを作成した．ここでラグの最大値は 500 とした．

生成したファイル間の NCD を求め，得られた距離行列を用いて MDS による分類およびクラスタ分析を行った．クラスタリングの手法として単連結法， 群平均法， 完全連結法を用いた．

3 結果

まずデータの種類における差異について述べる．ノイズの強さを示す定数 A を 0.1 および 1.0 に設定し，(a) 生データ，(b) 自己共分散関数のデータ，(c) 自己相関関数のデータに対して NCD を算出し，MDS による分類を行った．(a) の場合は分類できていないデータが存在したのに対し，(b) (c) の場合は概ね関数毎に分類された．ノイズの強い場合 ($A = 1.0$) では (b) のデータを用いると最もよく分類できた．またいずれの場合も $\sin(t)$ と $\cos(t)$ の区別は MDS では難しかった．

これら (a) (b) (c) のデータに対してクラスタ分析を行ったところ，(b) (c) のデータでは $\sin(t)$ と $\cos(t)$ を分類できた．本来一つになるべきクラスタが分裂していたり，一部のデータの位置が正しくない場合も生じたが，全般的には概ねうまく分類できていた．クラスター分析では (b) のデータを使用した場合に最もよく分類できた．

4 まとめ

本研究ではガウス型白色ノイズを有するデータに対して，正規圧縮距離を用いた分類が可能であるか調べた．生データのみならず自己共分散関数や自己相関関数を記録したファイルを用い，多次元尺度構成法およびクラスター分析により分類を行った．

この結果，生データはやや分類しにくいものの，自己共分散などの量に対して正規圧縮距離を適用することで概ねデータを分類できることがわかった．

一般に正規圧縮距離はノイズの影響を強く受けるが，本結果はノイズを有するデータであっても正規圧縮距離により分類できる可能性を示唆している．とりわけ自己共分散などの適切な量を用いることで，正規圧縮距離によりデータを分類できることがわかった．

参考文献

[1] R. Cilibrasi and P. Vitányi: 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] R. Cilibrasi, P. Vitányi and R. de Wolf: 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] 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).

[6] 石原 正道: 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, 第 72 回情報処理学会全国大会講演論文集, 1-255-1-256 (2010).

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

[8] 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).

[9] 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).