コルモゴロフ複雑度を用いた流れの不規則さ表現（不規則さの定量化）

Randomness Representation in Fluid Flows with Kolmogorov Complexity (Quantification of Randomness)

正 一宮 昌司（徳島大）

Masashi ICHIMIYA, University of Tokushima, 2-1 Minami-Josanjima, Tokushima, Tokushima, 770-8506 Japan

Key Words: Turbulence, Randomness, Kolmogorov Complexity, Normalized Compression Distance

1. 緒論

本研究は乱流の複雑さの問題を、最近各分野で研究されている「対象のコルモゴロフ複雑度」の観点から調べたものである。まずコルモゴロフ複雑度を紹介した後、その具体例である近似コルモゴロフ複雑度 AK や正規化圧縮距離 NCD を紹介し、 NCD に及ぼす圧縮形式やデータ個数の影響を検討する。これに続く講演では、実際に熱線データの複雑さを求めた結果を紹介する。

2. コルモゴロフ複雑度による複雑さ表現の原理

2.1 コルモゴロフ複雑度と圧縮度

コルモゴロフは情報理論と計算理論を結びつけ、有限列の不規則さ、ランダムネスをコンピュータで扱うことを考え、有限の離散列 X のコルモゴロフ複雑度 $K(X)$ を、対象 X を出力するプログラムの中で最も短いプログラムの長さとして定義した⁽¹⁾⁻⁽⁴⁾。Vitányi はさらに、与えられた文字列 X に対し、 X を復元可能な範囲で究極に圧縮したときに得られる二進列のビット長であると $K(X)$ を表現した⁽⁵⁾。

2.2 相対コルモゴロフ複雑度とデータ圧縮

真の $K(X)$ は計算不可能であることが証明されている⁽³⁾。このためランダムネスの純理論には有効な量であるが、実際のデータに利用することは考え難い。この点を克服するために、 $K(X)$ を近似する手法として性能の良い実用圧縮プログラム C によって得られるデータ X の圧縮列のビット長 $C(X)$ を $K(X)$ の近似として使う方法が提案された⁽⁵⁾⁽⁶⁾。

本研究は、実験データ（熱線風速計出力を A/D 変換した離散データ列）を圧縮プログラム C で調べ、その複雑度を近似するのに、元データ X の長さ $|X|$ に対する圧縮された X の長さ $C(X)$ の比、式(1)を近似コルモゴロフ複雑度 AK と名づけた。

$$AK = C(X) / |X| \quad (1)$$

実際には、単独の複雑度よりも対象集合中の元どうしの複雑度の区別が重要である。区別の基本は集合に定義される距離である。コルモゴロフは、 Y という補助情報があるときの X を計算するプログラムの中で最短のものの長さ相対コルモゴロフ複雑度 $K(X|Y)$ を定義した。 $K(X) - K(X|Y)$ は Y 中に於ける X の最大情報量であること、小さな誤差を無視すれば、 $K(X) - K(X|Y) \sim K(Y) - K(Y|X)$ であることが示されており、 X と Y の類似部分の量とみなすことが出来る⁽⁵⁾。これを用いて、データ間の情報距離 $E(X, Y) = \max\{K(X|Y), K(Y|X)\}$ が定義され、これが付加定数を除いて一般の距離 $d(X, Y)$ が満たすべき公理（同一性、対称性、非負性、3 角不等式(凸性)）を満たすことが示されている。この $E(X, Y)$ を基本として情報距離の正規化量 $NID(\text{Normalized Information Distance}) = E(X, Y) / \max\{K(X), K(Y)\}$ が考えられた⁽⁶⁾。

2.3 データ圧縮による対象間の正規化圧縮距離 NCD

$E(X, Y)$ を正規化した NID は理論的にはデータの違いの判

別量として適切であるが、 $K(X)$ が計算不可能であるから、やはり計算できない⁽⁶⁾。そこで $K(X)$ の近似 $C(X)$ を用いることに対応して正規化圧縮距離 $NCD(\text{Normalized Compression Distance})$ が提案されている。対象 X と Y の正規化圧縮距離 NCD を次式で定義する⁽²⁾⁽⁶⁾。

$$NCD(X, Y) = \frac{C(XY) - \min\{C(X), C(Y)\}}{\max\{C(X), C(Y)\}} \quad (2)$$

NCD を利用すると実際のデータのクラスタリングが可能で、言語、小説、画像、検索エンジンなどに適用されている。

3. NCD に及ぼす各影響

3.1 圧縮作業

本章では NCD を求める具体的な方法を説明する。圧縮に用いるデータ X, Y は、変動速度の実験データを数値の形に整えた数値データのテキストファイルと、画像データのビットマップファイルである。データ XY は、数値データの場合はデータ X の末尾にデータ Y の先頭が来るように繋いだ。このためデータ XY のファイル容量はデータ X のそれとデータ Y のその和になる。また画像データでは、描画ソフトにおいてデータ X の画像の上側にデータ Y の画像が空白なく位置するように繋いだものである。

圧縮作業では windows パソコンにおいて各ファイルを圧縮プログラムを用いて圧縮する。圧縮前のファイル容量 $|X|$ や圧縮後のファイル容量 $C(X)$, $C(Y)$ 及び $C(XY)$ を、式(1)や(2)に代入して $AK(X)$ や $NCD(X, Y)$ を求める。

3.2 圧縮形式の選定

圧縮形式にはさまざまなものが存在するが、 NCD に最適な圧縮形式について考察する。ここでは広く用いられている 7z, RAR, bzip2, LZH, ZIP, gzip の計 6 形式を試した。

数値データは二次元ノズルから噴出された噴流内で熱線風速計で測定された流れ方向変動速度 u を用いた。各データからその最小値を引き、これが 1 を越えないように、全測定データ間で共通の定数で正規化した後に、3 桁の整数にさせた。このような整数数値が一列に 98304 個（速度データは 32768 個）並んでおり、容量 98304 バイトのテキストファイルである。画像データは無次元化された流れ方向および垂直方向変動速度 262144 個の点を平面上にそれぞれ -1 から 1 までの範囲で散布した横幅 400 ピクセル、縦幅 400 ピクセルのビットマップファイルであり、容量は 480054 バイトである。

圧縮形式は距離公理の同一性すなわち $NCD(X, X) = 0$ をできるだけ満足する圧縮形式を選んだ。データ X には層流と乱流を代表するものを選んだ。表 1 に同じデータどうしの NCD を示す。7z 形式が最も 0 に近い。すなわち距離公理の同一性が最も成立する圧縮形式は 7z である。そこで 7z を以後の解析にはすべて用いることにした。距離公理の残りを 7z で確

Table 1 NCD between same data

data format	numerical		graphic	
	laminar	turbulent	laminar	turbulent
7z	0.013	0.003	0.315	0.062
RAR	0.045	0.007	0.635	0.969
bzip2	0.355	0.381	0.053	0.288
LZH	0.973	0.992	0.904	0.967
ZIP	0.979	0.975	0.703	0.971
gzip	0.969	0.957	0.906	0.990

Fig. 1 Distribution of NCD between same data as a function of data length.

認したが、対称性すなわち $NCD(X, Y) = NCD(Y, X)$ 、非負性すなわち $NCD(X, Y) \geq 0$ は計算した全データについて成立し、3角不等式すなわち $NCD(X, Y) + NCD(Y, Z) \geq NCD(X, Z)$ も多くの例で計算したところ成立した。

3.3 数値データ個数の影響

次に数値データ長さすなわち数値個数が NCD に及ぼす影響を調べ、適切なデータ長さを決定した。

まず前節と同様に、距離公理の同一性を調べるために同一数値データどうしの $NCD(X, X)$ を求めた。データ X には、786432 個の整数データの最初から n 個の整数を用い、 n を 384 から 786432 まで 12 通りに変えて NCD を計算した。図 1 に NCD の n に対する変化を示す。層流間の $n < 6 \times 10^3$ を除くと、 n が多いほど NCD は 0 に近づく。図 1 の結果からは、 n は多い方が望ましい。

次に距離公理の対称性を調べるために、 $NCD(X, Y) - NCD(Y, X)$ の n に対する変化を図 2 に示した。基準データ X には層流を、データ Y には層流と乱流を用いた。層流のデータ Y は X のごく近くのデータである。 n が小さい間は n によって値がかなり変化するが、 n が増加すると次第に 0 に近づき、距離公理の対称性が成り立つ。図 2 の結果からも、 n は多い方が望ましい。

次に、実際に NCD を求めるのは異なるデータ X, Y 間においてがほとんどなので、このときの $NCD(X, Y)$ の n に対する変化を求めた。図 3 に NCD の n に対する変化を示す。データ Y の層流データと乱流データは図 2 と同じである。 Y が層流でも乱流でも n が増加するに従って NCD は増加する。乱流では $n > 2 \times 10^4$ では NCD がほぼ 1 の一定値になる。またデータ Y は乱流の方が、層流である X とのランダムさとの差異が大きいのので NCD が大きい。NCD を用いて層流と乱流の区別を行う観点からは、 Y が層流のときの NCD と乱流のときのその差異が大きいために、 n は少ない方が望ましい。しかし n によって NCD が異なるような $n < 2 \times 10^4$ は望ましくない。

以上、図 1～3 の結果から、 $n = 98304$ (速度データ数 = 32768) を用いることに決定した。この速度データ数 32768

Fig. 2 Difference of $NCD(X, Y)$ and $NCD(Y, X)$ as a function of data length.

Fig. 3 Distribution of NCD between different data as a function of data length.

は、実際の AD 変換の取得時間では約 6.6 秒間に相当する。この n においては図 3 からわかるように、NCD は基準位置と層流間では 0.9 を越えており、一方基準位置と乱流間では約 1.0 となる。このように、NCD は Y が層流時と乱流時とで約 0.1 の範囲で変化する条件で考察を行うことになる。なお Y に基準データ X と同じデータを用いた場合には、表 1 や図 1 に示されているように $NCD = 0.013$ であるが、 Y として X のごく近くの層流を用いた場合には図 3 に示すように $NCD = 0.921$ となり、同じ層流間の X と Y であっても、データが異なるだけで NCD は 0.9 を越えてしまう。

4. 結論

有限離散列のランダムさを表現する測度であるコルモゴロフ複雑度とその近似について述べた。流れのランダムネスの定量的評価を行うために、熱線風速計出力の A/D 変換結果の有限離散列やコンピュータ上の 2 次元画像の有限離散列に、これを応用することを提案し、適切な圧縮形式やデータ個数を検討し、以下の結論を得た。

(1) 本研究の目的には、圧縮形式は調べた 6 種類の範囲では 7z が最適である。

(2) 熱線風速計出力数値データの NCD はデータ個数によって変化し、データ個数が多いほど同一データ間の NCD は 0 に近づき、異なるデータ間の NCD は 1 に近づく。

本研究には名古屋大学、中村育雄名誉教授のご指導を賜った。厚くお礼申し上げる。

文献

- (1) Li, M., and Vitányi, P. M. B., 渡辺治訳, コンピュータ基礎理論ハンドブック I (1994), p. 187, 丸善.
- (2) Li, M., and Vitányi, P. M. B., An Introduction to Kolmogorov Complexity and its Applications, 2nd ed. (1997), Springer.
- (3) 渡辺治, 数理科学, No. 519 (2006), pp. 10-15.
- (4) 渡辺治, 統計数理, Vol. 54, No. 2 (2006), pp. 511-523.
- (5) Vitányi, P., 渡辺治訳, 数理科学, No. 519 (2006), pp. 54-59.
- (6) Li, M., Chen, X., Li, X., Ma, B., and Vitányi, P.M.B., IEEE Trans. Inf. Theo., Vol. 50, No. 12 (2004), pp. 3250-3264.