コルモゴロフ複雑性に基づく画像圧縮と分類に関する実験と考察
Image compression and clustering based on Kolmogorov complexity

齋藤 高央 † 湊 真一 † ツォイクマン トーマス †
Takao Saitoh Shin-ichi Minato Thomas Zeugmann

1. まえがき

近年、インターネットなどの普及により、多量の画像データを扱う機会が増えてきている。多量の画像データを手作業で分類するのは困難であり、自動的に分類できる手法へのニーズが高まりつつある。

情報の複雑さを表す指標として、コルモゴロフ複雑性が知られている。[5] コルモゴロフ複雑性に基づく類似度判定手法の一つとして、圧縮度を用いた手法が提案されている。[3]

本研究では、パラメータフリーな画像の分類手法の確立を目指している。そこで基本的性質を調査するために、まず画像に対してコルモゴロフ複雑性に基づく圧縮度を用いた類似度判定を行い、画像を分類する実験を行った。

2. コルモゴロフ複雑性と正規圧縮距離

コルモゴロフ複雑性は、有限長のデータ列の複雑度を表す指標の一つで、与えられた文字列 x のコルモゴロフ複雑性 $K(x)$ は、$K(x)$ を出力するプログラムの最短長で表される。

このコルモゴロフ複雑性を用いたデータ間の距離の定義として、Normalized Information Distance(以下 NID) が提案 [1][2][6] されている。文字列 x, y 間の距離 $NID(x, y)$ は、以下のように定義される。

$$NID(x,y) = \frac{Max(K(x), K(y))}{Max(K(x|y), K(y|x))}$$

ここで、$K(x|y)$ は、y を入力としたときに x を出力するプログラムの最短の長さであり、$K(x)$ は、入力なしで x を出力するプログラムの最短の長さである。しかし、$K(x)$ は原理的に計算不能であることが知られており [5]、$NID(x, y)$ を直接求めることができない。

そこで、コルモゴロフ複雑性を近似的に求める手法が提案 [3] されている。具体的には、ある文字列 x をある圧縮プログラム C を用いて復号可能な状態で究極に圧縮した時のビット長を $C(x)$ と表し、これを文字列 x のコロモゴロフ複雑性 $K(x)$ の近似値と考える。

この近似を用いたデータ間の距離を Normalized Compression Distance(以下 NCD) と呼び、下のように定義される。

$$NCD(x,y) = \frac{C(xy) - Min(C(x), C(y))}{Max(C(x), C(y))}$$

ここで、$C(x)$ はある文字列 x を圧縮プログラム C を用いて圧縮した時のビット長であり、$C(xy)$ は文字列 x, y を結合したものを圧縮プログラム C を用いて圧縮した時のビット長である。

†北海道大学 大学院 情報科学研究科

本実験では、画像のみを扱うこととしたので、圧縮形式も画像特有の形式を用いた方がよいと判断し、可逆圧縮であり、扱える色数が多い PNG 形式を用いた。

3. 実験

本実験は、Pentium M 1.80GHz、RAM 2GB の計算機で、OS は WindowsXP SP3 を用いて行った。実験プログラムは C#を用いて作成し、画像の圧縮には ImageMagick の C#向けライブラリである MagickNet を用いた。

3.1 概要

本実験の目的は、基本的な性質を把握することであるので、比較的単純なパターンの画像を用いて行うこととした。分類結果の判定がしやすいものとして、10 カ国の国旗 (図 1) を用い、以下の手順で実験を行った。

図 1: 実験に用いた 10 カ国の国旗

まず前処理として、画像を縦横比 4:3 に揃えた 24bit の Microsoft Windows Bitmap Image 形式に変換した。これは、現在の結合処理の都合で、画像の縦横比を揃えておく必要があったからである。

次に画像を結合・圧縮して距離を計算する。これをすべての組み合わせに対して繰り返し実行し、距離行列を作成した。この距離行列を基にクラスタリングを行い、結果を比較した。

画像ファイルを扱う場合、単純に上下に画像を結合すると、元の画像で同じ位置にあった部位が離れて配置されてしまうことになる。例えば、元の画像で左上だった部分は、結合後の画像では左上と中央付近に配置されることになる。

これを改善するため、画像を 1px 幅でインターリーブ結合する方式での実験を行った。インターリーブ結合の場合、元画像で左上端にあったピクセル同士が、結合後の画像でも左上端で上下に並ぶことになり、より圧縮率の改善が見込めると考えたためである

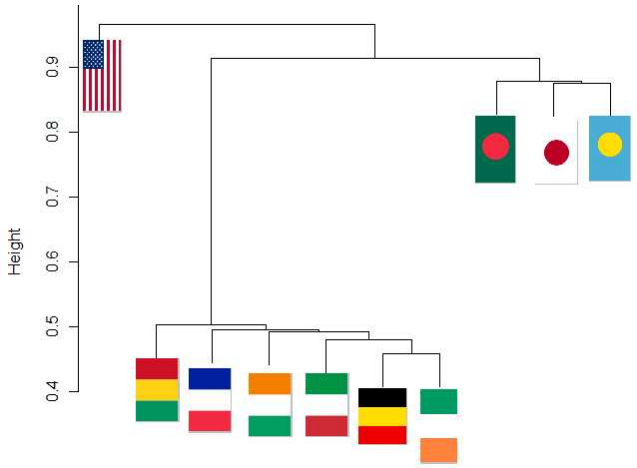

図 2: 本手法を用いた距離行列によるクラスタリング結果

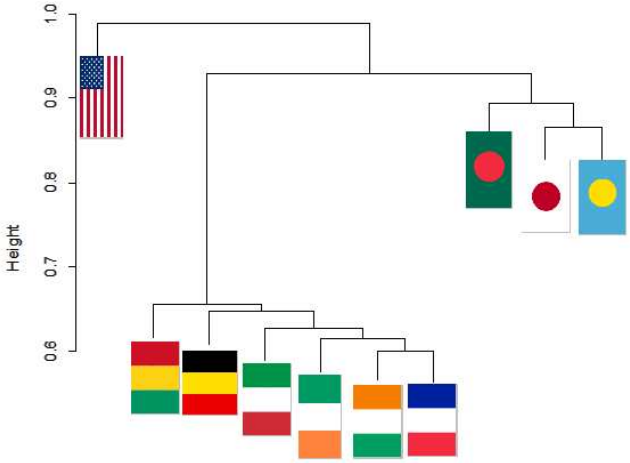

図 3: Complearn を用いた距離行列によるクラスタリング結果

4. 結果

4.1 Complearn との比較

圧縮度ベースの距離を計算するツールとしてよく知られているものに Complearn[3] がある。Complearn は汎用の圧縮アルゴリズムを用いて距離を計算するツールであるので、汎用アルゴリズムと画像向けのアルゴリズムの結果を比較することができる。比較に用いたのは Complearn バージョン 0.9.7 で、圧縮形式は zlib を用いた。

Complearn はクラスタリングを行う機能を持っているが、本実験においてはクラスタリング手法を統一する必要があったので、この機能は使用していない。

クラスタリングを行った結果は図 2,3 のようになる。両者とも星条旗、縦縞を図案に持つ国旗、丸を図案に持つ国旗の 3 つのクラスタができていることが分かる。縦縞のクラスタの距離に一部差があるが、全体としてみると類似した結果となった。PNG 形式は内部的に zlib を使っているので、この結果は妥当と考えられる。

4.2 結合方式による比較

単純な上下結合とインターリーブ結合での結果 (図 4) を比較した。単純な画像の場合、インターリーブ結合を行ったことで結合時の圧縮率が悪化し、クラスタリング結果に悪影響を与えていると推測できる。

4.3 自分自身との距離

距離行列の対角線は自身との距離であるので、0 になる事が望ましい。実験の結果得られた距離行列は表 1 のようになった。これは我々が利用することのできる圧縮プログラムが理想的な動作をする圧縮プログラムではなく、理想的な圧縮結果を得ることはできないためであると考えられる。自身との距離は近似的に 0 とみなす事とするが、ややばらつきが大きい。

例えば、アメリカ合衆国の場合は自身との距離が 0.0559 となり 0 に近似できるが、フランスの場合は 0.3713 となり、0 に近似するにはやや大きい。Complearn

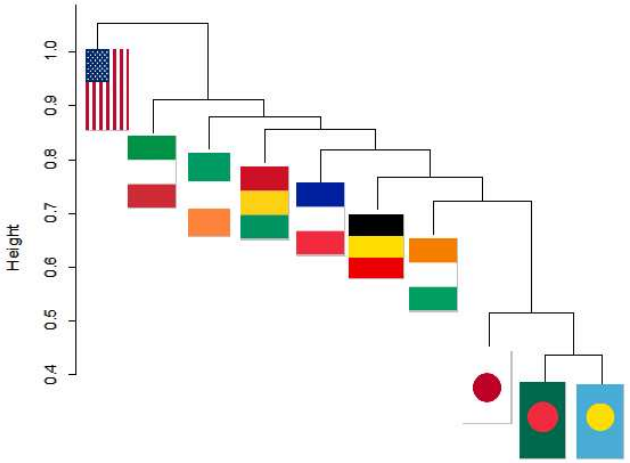

図 4: 本手法を用いたインターリーブ結合でのクラスタリング結果

を用いた場合の距離行列も表 2 のようになり、類似した傾向にある。

そこで、圧縮後のファイルサイズを調べてみると、表 3 のようになった。圧縮後のファイルサイズが小さいものほど、自分自身との距離が大きくなっていることがわかる。この傾向は同じクラスタに分類された他の国旗の場合でも共通しており、圧縮後のファイルサイズは（縦縞の図案を持つ国旗）＜（丸の図案を持つ国旗）＜（星条旗）となっていた。

これらの結果から、圧縮されやすい画像の場合、結合時の圧縮率が改善されず近似が成立しづらくなっていると考えられる。

表 1: 実験プログラムによる距離行列（一部抜粋）

国名	バングラデシュ	フランス	アメリカ
バングラデシュ	0.1537	0.9192	0.9593
フランス	0.9167	0.3713	0.9740
アメリカ	0.9600	0.9733	0.0559

表 2: Complearn による距離行列（一部抜粋）

国名	バングラデシュ	フランス	アメリカ
バングラデシュ	0.1858	0.9354	0.9806
フランス	0.9354	0.4051	0.9946
アメリカ	0.9806	0.9942	0.0592

表 3: 圧縮後のファイルサイズ

	バングラデシュ	フランス	アメリカ
単独時 (byte)	1177	237	3077
結合時 (byte)	1293	312	3341

5. 終わりに

汎用の圧縮形式に変えて、画像特有の圧縮形式を用いて NCD を計算する手法に関して実験を行った。本実験では、内部的に汎用の圧縮形式に近いアルゴリズムを使う PNG 形式を選択したため、汎用の圧縮形式との違いがあまりない結果となった。画像向けの圧縮形式を使用した場合でも、既存の形式と同程度の分類結果を得られることが分かった。

本実験においては単純な上下方向の画像結合が有効であったが、風景写真を用いて行った実験ではインターリーブ結合の方が優勢な結果が得られるなど、現段階ではすべての種類の画像をパラメータフリーで分類する事はできていない。また、類似の色成分を集中させるためにブロックソートなどを用いる手法も考考えられる。

今後は、画像特有の成分を活用する可逆圧縮形式などを用いての実験を行う予定である。

謝辞

本研究の一部は、文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」（領域番号 456 による）

参考文献

[1] Charles. H. Bennett, Peter Gacs, Ming Li, Paul M. B. Vitányi and W.H.Zurek: “Information Distance,” IEEE Transactions on Information Theory, 44, 4, pp.1407-1423 (1998)

[2] Rudi Cilibrasi and Paul M. B. Vitányi: “Clustering by compression,” IEEE Transactions on Information Theory, 51, 4, pp.1523-1545(2005)

[3] Rudi Cilibrasi and Paul M. B. Vitányi: “Similarity of objects and meaning of the words,” Theory and Applications of Models of Computation, Third International Conference, TAMC 2006, Beijing, China, May 2006, Proceedings, Vol.3959 of Lecture Notes in Computer Science, Berlin Springer, pp.21-45 (2006) (Complearn http://www.complearn.org/)

[4] Robert Gentleman and Ross Ihaka etal. “R Documentation http://www.r-project.org/other-docs.html”

[5] A.N. Kolmogorov “Three approaches to the quantitative definition of information,” Problems Inform. Transmission, 1:1(1965), 1-7.

[6] Ming Li, Xin Chen, Xing Li, Bin Ma and Paul M. B. Vitányi: “The similarity metric,” Proc. 14th ACM-SIAM Symposium on Discrete Algorithms (SODA), pp.863-872 (2003)