Kolmogorov 記述量に基づく類似度を用いた方言の自動分類

Automatic classification of Japanese dialects using similarity metric is based on Kolmogorov Complexity

http://www.tani.cs.chs.nihon-u.ac.jp/ midori/

谷　研究室　　大江　碧

Midori Oe

概要

Kolmogorov 記述量に基づいたデータ間の類似度に関する距離が定義され，ＤＮＡの類似度や言語の類似度，音楽の類似度判定に有用だという実験結果が得られている．本研究では Kolmogorov 記述量が方言の類似度分析に対して有用かどうか実験を行う．

1　はじめに

常に全ての事柄は何かによって分類されている. 今日，コンピュータ技術の発達に伴い，文章や音楽も「データ」として扱われるようになってきた．Ray J.Solomonoff, Anderei Kolmogorov, Gregory J.Chaitin らは，Kolmogorov 記述量に基づくデータ間の類似度に関する距離を表す similarity metric を発案した．これを用いて，ＤＮＡの類似度や言語の類似度，音楽の類似度判定に有用だという実験結果が得られている．

これまで方言は，単語単位では類似度解析がされてきたが，文章を対象とした類似度解析はされたことがなかった．そこで本研究では,「方言ももたろう」（株式会社富士通ビー・エス・シー）というソフトのデータを対象として類似度解析を行っていく．

第 2 節では Kolmogorov 記述量について，第 3 節では similarity metric について述べ，第 4 節では実験の概要としてデータや結果の表示方法について解説し，第 5 節では実験結果・考察を述べ，まとめとする．

2　Kolmogorov 記述量の定義

この章では Kolmogorov 記述量の基本的な定義について述べておく．

2.1　Kolmogorov 記述量

あるデータ x があるとする．データ x の Kolmogorov 記述量とは，直感的には，あるプログラミング言語で x を生成する最小のプログラムサイズ（すなわちビット数）のことである．ただし，この場合プログラムは，空の状態からスタートし，すべてを出力するものとする．どんなプログラミング言語を選んでも，それが妥当であれば，情報量は定数分の差しかないことが証明されている．

S によるデータ x の Kolmogorov 記述量 $K_s(x)$ は以下のように定義される．

$$K_s(x) = min|p| : S(p) = x$$

このとき，S をプログラム言語，$|p|$ をプログラムサイズとする．つまり上記の式は，プログラミング言語 S で記述した x を生成するプログラムのうちサイズが最小のもののサイズのことである．

また，あるデータ y が与えられているときのデータ x の記述量を定義する．

また，x と y を区別できる形で出力させる最短のプログラム長を，$K(xy)$ と表す．$O(logK(xy))$ の誤差範囲では，

$$K(xy) = K(x) + K(y|x)$$

となることが証明されている．

2.2　K のアルゴリズム的性質

$K(x)$ を正の整数から正の整数への関数と考える．

定理

1. 関数 $K(x)$ は帰納的ではない．しかも，どのような帰納的部分関数 $\phi(x)$ を考えても，もしその値が無限個の点で定義されているならば，その定義上のどこかの点で $\phi(x) \neq K(x)$ となる．

2. 引数 t に対して単調減少で (全域的な) 帰納的関数 $H(t,x)$ が存在し，$\lim_{t\to\infty} H(t,x) = K(x)$ ．すなわち，$K(x)$ のよい近似は計算可能．ただしこれは一様近似ではない．

2.3 情報量

相対記述量 $K(x|y)$ が絶対記述量 $K(x)$ よりかなり小さい場合,「y が x についての情報を多分に含んでいる」と考えることができる．したがって，定数分の差異を無視すれば,“ y に含まれる x の情報量 ”は

$$I(x:y) = K(y) - K(y|x)$$

とみなすことができる．また，$K(x|x)$ となる x を選べば,

$$I(x:x) = K(x)$$

となる．ここで考えるべきことは，情報の対称性である．

$$OlogK(xy)$$

の範囲では,

$$K(xy) = K(x) + K(y|x)$$

が成り立つので，これを用いると,

$$I(x:y) = I(y:x)$$

が成り立つ.

3 similarity metric

数学において距離空間とは，任意の 2 点間で距離が定められた空間のことをいう.

定義

ある集合 X 上の距離とは，実数値関数 $d : X \times X \to R$ で任意の $s, y, z \in X$ に対して次のような性質を満たす.

$$d(x,y) \geq 0$$
$$d(x,y) = 0 \Leftrightarrow x = y$$
$$d(x,y) = d(y,x)$$
$$d(x,y) \leq d(x,z) + d(z,y) : \text{三角不等式}$$

これをもとに，情報距離について考える.

Ming Li らの研究では，情報に関する距離を標準化している．任意の文字列 x, y について，以下のように決める.

$$d(x,y) = \frac{max\{K(x|y), K(y|x)\}}{max\{K(x), K(y)\}}$$

また，$K(y) \geq K(x)$ としたとき,

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

これに対して先に述べた情報量の公式を用いると,

$$d(x,y) = \frac{K(y) - I(x:y)}{K(y)}$$

となる．また 2 節で示した通り，$O(\log K(xy))$ の範囲では $K(xy) = K(x) + K(y|x)$ が成り立つので,

$$d(x,y) = \frac{K(xy) - K(x)}{K(y)}$$

と表すことができる.

実際の実験では，この式とともに以下の 3 つの理想理論のもとに行われた.

(1) 要求された情報距離 $d(x,y)$ は漠然と長い文字列 x, y によって得られる.

(2) Kolmogorov Complexity は帰納的でないため，計算不可能である.

(3) 実用的な方法で情報距離を近似する際，圧縮方法の一つである“ bzip2 ”を用いる.

以上より、$bzip(x)$ を文字列 x を bzip2 で圧縮したときのファイルサイズとすると情報距離 $d(x,y)$ は以下のように近似される.

$$d(x,y) \fallingdotseq \frac{bzip(xy) - bzip(x)}{bzip(y)}$$

この距離関数をもとに，データの分類を進めていく.

4 実験の概要

4.1 対象データ

今回の実験で取り扱ったデータは「方言ももたろう」（株式会社富士通ビー・エス・シー）というソフトに入っている WAV ファイル，WAV ファイルをテキストファイルにしたもの，WAV ファイルを mp3 ファイルに変換したものである.

音声ファイル（WAV ファイル，mp3 ファイル）は bzip2 で圧縮し、第 3 節で述べた

$$d(x,y) \fallingdotseq \frac{bzip(xy) - bzip(x)}{bzip(y)}$$

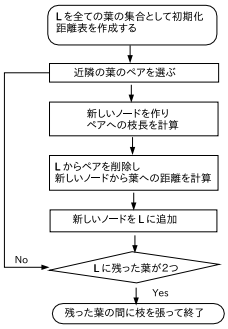
に当てはめる.

テキストファイルは bzip2，compress，zip，gzip，lzh で圧縮し上記の式に当てはめる.

4.2 NJ 法

無根系統樹作成法の一種である．NJ 法は個々の系統の進化速度が異なっているときに特に有効である．(Saitou and Nei 1987)

4.3 NJ 法のアルゴリズム



4.4 STEP1

- L を全ての葉の集合として初期化
- $r(i)$ をもとめる

$$r(i) = \sum_{k \in L, k \neq i}^{N-1} d(i,k)$$

4.5 STEP2

- M_{ij} を計算

$$M(i,j) = d(i,j) - \frac{[r(i) + r(j)]}{N-2}$$

N：葉の数 (データ数)

4.6 STEP3

- M_{ij} が最小となるペア i，j を選ぶ
- 新しい節点 l を作る
- 節点 l から i，j への枝長を計算

$$S(il) = \frac{d(ij)}{2} + \frac{[r(i) - r(j)]}{2(N-2)}$$
$$S(jl) = d(ij) - S(il)$$

4.7 STEP4

- l から i，j 以外の L に含まれる node への距離 d を計算する

$$q \in L, q \neq i, j$$
$$d(ql) = \frac{d(iq) + d(jq) - d(ij)}{2}$$

- L から i，j を除き，l を追加

$$N = N - 1$$

4.8 STEP5

- if $(N = 2) \rightarrow$ 終了
- else $\rightarrow$ STEP1 へ

5 実験結果・考察

まず WAV ファイルを bzip2 で圧縮したもの，テキストファイルを zip，gzip，lzh で圧縮したものは Kolmogorov 記述量に基づく距離

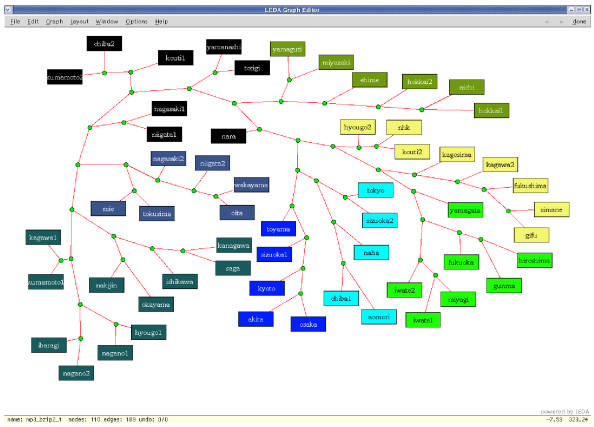
$$0 <= d(x,y) <= 1$$

を満さなかったため，可視化する対象からはずした．

5.1 実験結果

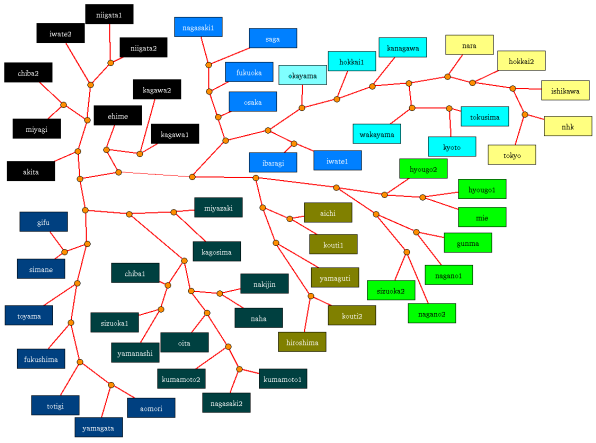
mp3

WAV ファイルを mp3 ファイルに変換し，bzip2 で圧縮したときの図



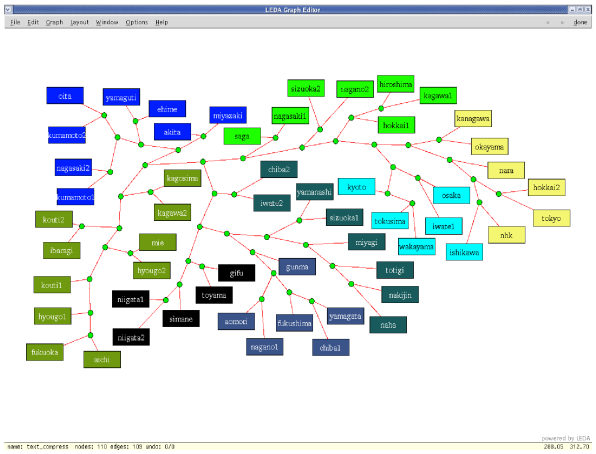
bzip2

WAV ファイルを text ファイルに書き起こし，bzip2 で圧縮したときの図



compress

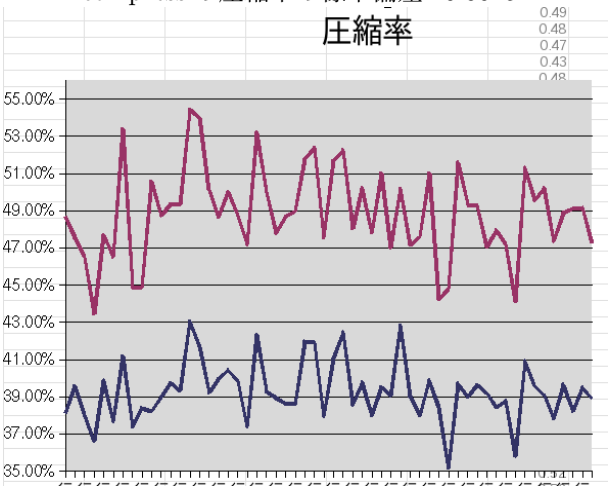
WAV ファイルを text ファイルに書き起こし，compress で圧縮したときの図



bzip2 と compress の圧縮率の違い

bzip2 の圧縮率の標準偏差=0.0006

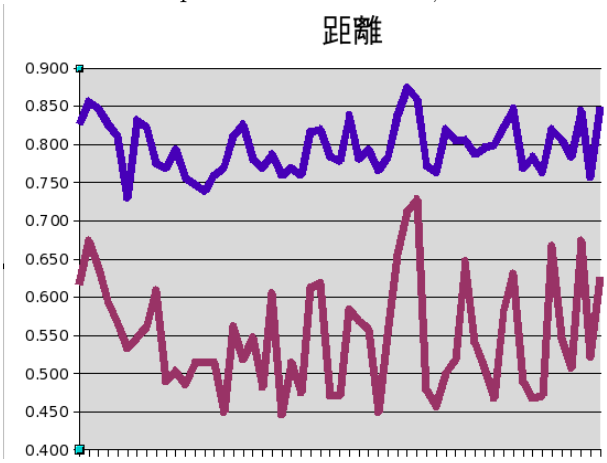
compress の圧縮率の標準偏差=0.00254



bzip2 と compress の距離の違い

bzip2 の距離の分散=0.009

compress の距離の分散=0,002



5.2 考察

bzip2 と compress の違いがグラフから読み取れるが，どちらが Kolmogorov 記述量に基づく距離を算出するのに有用であるかはわからなかった．先行研究では bzip2 が Kolmogorov 記述量に基づく距離を算出するのにいいのではないかといわれている．

また，音声ファイルについては，今回はテキストファイルと比較すると類似性があまりでなかったように思える．WAV や mp3 以外の形式や，bzip2 以外の圧縮方法を試してみることも必要ではないかと思う．

参考文献

[1] 谷 聖一　データの複雑さ・データ表現の複雑さ

[2] 三井 貴子　Kolmogorov Complexity を用いた民族音楽分類，2005

[3] 西村 久香　系統樹作成アルゴリズム性能評価実験，2005