

ライス符号を用いた方言の符号化について

Automated classification of dialect that uses rice code

谷 研究室 益田 真太郎

Shintaro Masuda

概要

圧縮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であるという実験結果が得られている。近年、谷研究室では圧縮類似度を用いた方言の自動分類を行ってきた。本研究では方言データにおいて、文字を数値に置き換えて単純化したデータをさらにライス符号化することを提案する。本提案が分類の精度の向上に有効かを確かめるため、具体的な実験を作った。

1 はじめに

現在様々な事柄がデータ化されている。もちろん、音楽や文章も「データ」として扱われるようになってきている。Ming Li らが圧縮類似度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 (現在 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ことが分かっている。) した。[1] また、圧縮類似度距離はデータが小さいほどより近似される。

谷研究室では、圧縮類似度を用いた方言の自動分類について研究している。先行研究では、方言の音声データをテキスト化し、そして、前処理を加えて数値化していた。[2] 本研究では、従来の前処理で数値化されたデータを符号化するという新たな前処理を行う。符号化するにあたり、実装が簡単で画像や音声の圧縮アルゴリズムの中でよく使われているライス符号を使うことにする。[3] 本研究では、ライス符号化によりデータが小さくなるか確かめるため、従来の前処理を施したデータと新たに従来の前処理を施したデータにライス符号化を施したデータとのファイルサイズ比較を行う。なお、前処理を施したデータは圧縮を施した場合と圧縮を施していない場合の 2 パターンでファイルサイズ比較を行う。圧縮方法に関しては、先行研究で良い結果が出ている bzip2 を使う。第 2 章では圧縮類似度について、第 3 章ではライス符号について、第 4 章では実験データ、実験概要、第 5 章では実験結果、考察、第 6 章では今後の課題を述べる。

2 圧縮類似度

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 $d: X \times X \rightarrow R$

で任意の $x, y, z \in X$ 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, Xin Chen らの研究では、情報に関する距離を標準化していて、任意の文字列 $x; y$ について、以下のように決めている。

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

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、 $O(\log K(xy))$ 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。

kolmogorov 記述量は計算でその値を算出することが不可能であるため、圧縮プログラムを用いて値を近似的に求めることになる。 $bz2(x)$ 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。

$$d(x, y) \doteq \frac{bz2(xy) - bz2(x)}{bz2(y)}$$

また圧縮方法 C が以下を満たすとき $O(\log n)$ の誤差の範囲で万能であると証明されている。

1. 任意の文字列 x に対して $C(xx) = C(x)$ であり、空文字列 λ に対して、 $C(\lambda) = 0$ である。
2. 任意の文字列 $x; y$ に対して $C(xy) \geq C(x)$ 。
3. 任意の文字列 $x; y$ に対して $C(xy) = C(yx)$ 。
4. 任意の文字列 $x; y; z$ に対して $C(xy) + C(z) \geq C(xz) + C(yz)$ 。

3 整数の符号化

整数の符号化とは、主にデータ圧縮において整数を符号語にするための手法である。多くの場合は任意の、あるいは決められた範囲内の整数をなるべく短い符号語で表現し、全体の情報量を少なくすることを指す。整数の符号化によって出力される符号語は一意復号可能であり、ふつう接頭符号 (prefix code) である。一意復号可能な符号語は、一般に語頭条件を満たす。そのため接頭符号を語頭符号とも呼ぶ。

3.1 アルファ符号

アルファ符号とは、一進法符号 (単進符号, unary) とも呼ばれる、正の整数を表す可変長符号の一つ。小さな数には短い符号語を、反対に大きな数には長い符号語を割り当てる。可変長符号の最も簡単な形であるが、符号語が長くなりすぎることからアルファ符号単体で用いられることは少ない。実際にはこの応用であるガンマ符号やデルタ符号が用いられる。

3.2 ライス符号

ゴロム (Golumb) 符号は Solomon W. Golumb が開発した符号である。ライス (Rice) 符号はゴロム符号の特別な場合で、Robert F. Rice がのちに再発見したことから、ゴロム・ライス符号と呼ばれている。ゴロム符号はパラメータ b を使って整数 n を符号化する。アルゴリズムは次のようになる。

1. 商 $p = n/b$, 剰余 $q = n \% b$ を求める。
 2. p をアルファ符号で符号化する。
 3. $b = 2^k$ の場合 q の k ビットをバイナリ符号化する。
 4. $b! = 2^k$ の場合、 q を CBT 符号で符号化する。
- 3 の場合をライス符号と言う。ライス符号はパラメータ b を変更すると符号語も変化する。今回の実験では、 $b = 8, 16$ の場合で実験を行う。

4 実験データ・実験概要

実験データ

方言ももたろう (監修・著: 杉藤美代子 (音声言語研究所所長)) というソフトに日本各地 56 箇所の音声データがある。その音声データをテキスト化 (ひらがな・。のみ使用) また「を」→「お」、「へ」→「え」、「は」→「わ」、伸ばす音は前の音の母音をもう一つ加えるという文章の書き方をした。

前処理 1(pre1)

50 音順で 1 から対応させて変換する。

前処理 2(pre2)

使用頻度が高い順に 1 から対応させて変換する。

前処理 1-r(pre1-r)

前処理 1 を施したデータにライス符号化を施す。ライス符号化は $b = 8, 16$ の 2 種類。

前処理 2-r(pre2-r)

前処理 2 を施したデータにライス符号化を施す。ライス符号化は $b = 8, 16$ の 2 種類。

実験概要

ライス符号化によりデータが小さくなるか確かめるため、前処理 1,2 を施したデータ (先行研究) と前処理 1-r,2-r を施すデータ (本研究) とのファイルサイズを比較する。

実験は、以下のように行う。

■圧縮 (bzip2) 前

前処理 1,2 を施したデータと前処理 1-r,2-r のを施したデータとのファイルサイズ比較。

ライス符号が $b = 8, 16$ の 2 種類で行う。

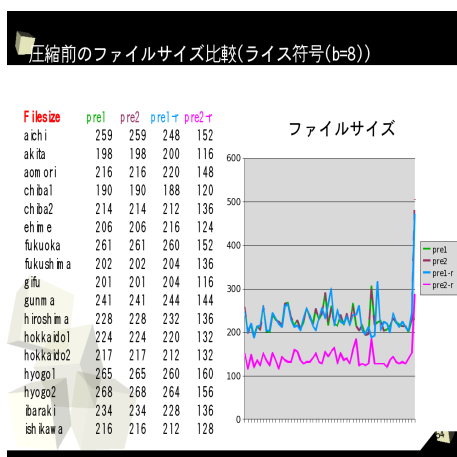
■圧縮 (bzip2) 後

前処理 1,2 を施したデータと前処理 1-r,2-r を施したデータとのファイルサイズ比較。

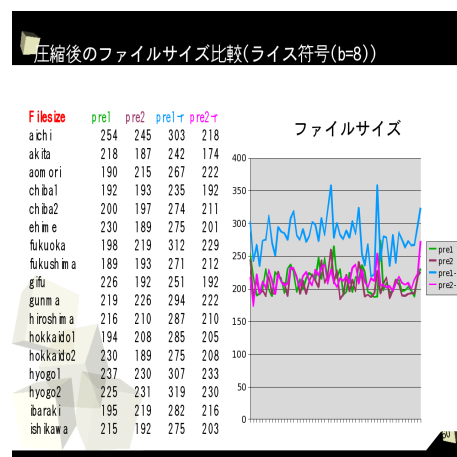
ライス符号が $b = 8, 16$ の 2 種類で行う。

5 実験結果・考察

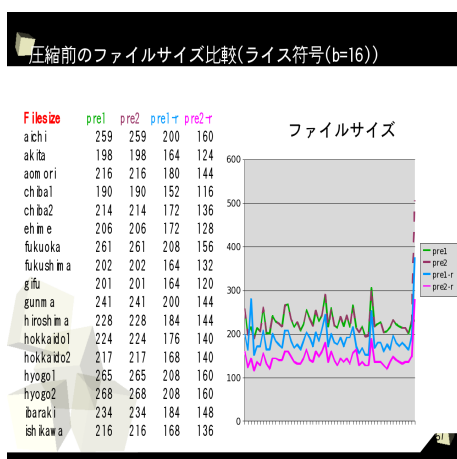
■圧縮 (bzip2) 前



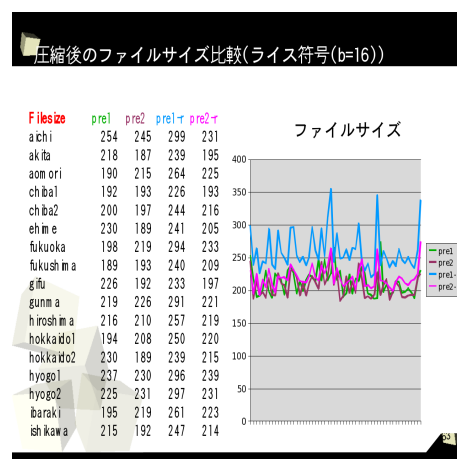
前処理 2-r を施したデータのファイルサイズが大幅に小さくなっている。



前処理 1-r を施したデータのファイルサイズが大幅に大きくなっている。



前処理 1-r,2-r を施したデータ共にファイルサイズが大幅に小さくなっている。



前処理 1-r を施したデータのファイルサイズが大幅に大きくなっている。

■圧縮 (bzip2) 後

6 今後の課題

ライス符号化が方言の分類に有効かどうかは現時点では、わからなかった。今後、bzip2 以外の圧縮方法で比較実験を行い、引き続き、ライス符号化が方言の分類に有効か実験を行う。

参考文献

- [1] Ming Li and Paul M.B.Vitanyi, 渡辺治翻訳 KolmogorovComplexity and its Applications. コンピュータ基礎理論ハンドブック (1994)
- [2] 堀中 幸司, Kolmogorov 記述量に基づく類似度を用いた方言の自動分類 (2007)
- [3] S. W. Golomb, Run-length encodings, IEEE Transactions on Information Theory IT-12(3), 399-401.
Robert F. Rice, Some Practical Universal Noiseless Coding Techniques, JPL Publication 79-22, 1979.