連結クラスタリング法を用いた方言の自動分類

Automated classification of dialect that uses connected clustering

谷　研究室　　佐藤　雄太
Sato Yuta

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。2007 年度、日本大学文理学部情報システム解析学科谷研究室、堀中、高野、関根が Kolmogorov 記述量を用いた方言の類似度分析において、NJ 法、UPGMA 法、Qualtet Method の３種類のクラスタリング手法を用いて方言の自動分類を行った。[2] 本研究では、日本電気（株）藤原、後藤、井口が提案した連結クラスタリング法が、方言の分類に有用かどうかを調べる実験を行う。

1　はじめに

現在の方言研究というものは単語単位、アクセント、イントネーションによる研究がなされている。しかし文章単位での研究というものは余り進められていない。そこで Ming Li らが 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（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。）により、2007 年度、日本大学文理学部情報システム解析学科谷研究室、堀中、高野、関根が Kolmogorv 記述量を用いた方言の類似度分析を行った。主に、文字コード、前処理、クラスタリング手法（NJ 法、UPGMA 法、Quartet Method　の３種類）の違いでの研究を行った結果、クラスタリング手法では、NJ 法の有用性を確かめることができた。[2] 本研究では、堀中らの実験で得られたデータを、日本電気（株）藤原、後藤、井口が提案した連結クラスタリング法 [1] を用いて方言の自動分類を行い、連結クラスタリング法の有用性を調べる実験を行う。

第２章では圧縮類似度について、第３章ではクラスタリングについて、第４章では実験概要、第５章では実験結果、第６章では考察を述べる。

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 $d : X \times X \to R$ 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 2 節で示した通り，$O(logK(xy))$ 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。

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

また圧縮方法 C が以下を満たすとき $O(logn)$ 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 $C(xy) \geqq C(x)$。
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 $C(xy) + C(z) \geqq C(xz) + C(yz)$。

3 クラスタリング

今回は NJ 法、UPGMA、quartet-method の 3 つのクラスタリング手法を採用した。 これらはどれだけ似ているかを示す尺度として距離を用いるという性質があることと、例年との比較を行うために使用した。

3.1 NJ 法

近隣結合法と呼ばれ、全ての枝の長さの総和が最小になるように系統樹を作成していく無根系統樹。 効率が良く、他の方法では扱えないような大量のデータを扱うことができるが、出力された木が最適とは限らない。

3.2 UPGMA 法

平均距離法と呼ばれる単純な系統樹制作法で、最も近縁なクラスタ間の距離の平均を求めながら系統樹を作成する。 葉から決まり、最後の根の位置が決まる有根系統樹。

3.3 Quartet Method

木が距離表にどの程度沿っているか評価する基準があり、ランダムに木を変更し評価値を更新していくヒューリスティック（試行錯誤）アルゴリズムである。 NJ 法、UPGMA 法と比較すると大幅に処理時間を要する。

3.4 連結クラスタリング法

連結クラスタリング法は、文書を連結しながらクラスタを併合する方法であり、図 1 の手順で実行する。まず、文書集合 D と求めるクラスタ数 K が入力されると、各文書を各クラスタとした要素数 N のクラスタ集合 C を作成する。次に、文書ペア間の非類似度 NCD を計算し、最も類似度の高い（すなわち NCD の小さい）文書ペア d_x と d_y を探す。

Algorithm： 連結クラスタリング法

input： D：文書集合 $\{d1, d2, \cdots, dN\}$
K：クラスタ数

output： C：クラスタ集合 $\{C1, C2, \cdots, CK\}$

Begin

$C = \{\{d_1\}, \{d_2\}, \cdots, \{d_N\}\}$
$Fork = 1, 2, \cdots, N - K$
$(d_x, d_y) = argminNCD(d_i, d_j)$
$d_x = d_x \cdot d_y$
$D = D|d_y$
$C_x = C_x \cup C_y$
$C = C|C_y$

End

図 1 連結クラスタリング法

それから，最も NCD の小さい文書ペアを連結した $d_x \cdot d_y$ を改めて d_x と、d_y を文書集合 D から削除する。ここで、文書ペアの連結とは、文書 d_x と d_y を順に並べた文書である。そして、クラスタ C_x と C_y を併合したクラスタを改めて C_x とし、クラスタ C_y を削除する。これら文書ペア選択，文書の連結，クラスタの併合という動作を繰り返すことにより、連結クラスタリング法は、文書をクラスタリングする．$N - K$ 回繰り返すとクラスタ数が K となるのでクラスタリングを終了し、クラスタ集合 C を出力する。

4 実験概要

全国 56 箇所を 9 地方に分割する。(北海道、東北、関東、中部、近畿、中国、四国、九州、沖縄)
そして、クラスタ集合の数を 10 とし、それぞれのクラスタ集合中のデータが、9 地方にどのように分かれているかを検証する。

実験データ

方言ももたろう（監修・著：杉藤美代子）というソフトに昔話「桃太郎」が日本各地５６箇所の方言で読まれた音声データがある。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。[2]

前処理

使用頻度が高い順に 1 から対応させて変換する。[2]

5　実験結果

どのクラスタ集合にどこの地域がいくつ分類されているかを表す。$c1$〜$c10$…クラスタ集合

- $c1$…北海道 2, 関東 4, 中部 3, 近畿 6, 中国 2, 四国 4
- $c2$…九州 2
- $c3$…東北 1, 四国 1, 九州 2
- $c4$…東北 1, 関東 4, 四国 2, 九州 4
- $c5$…東北 2
- $c6$…中部 1, 九州 1
- $c7$…東北 2, 関東 4, 中部 2, 中国 1
- $c8$…中部 1, 九州 1
- $c9$…沖縄 1
- $c10$…沖縄 1

6　考察

局所に地域ごとのまとまりも見られるが、全体的には各地方がばらばらになった。今回は実験に大幅な時間を要したので、前処理を施したデータを 1 種類しか実験できなかった。今後は他の前処理を施したデータを用いて実験することが課題。そして、他 3 つのクラスタリング手法と比較するために、最終的に ”木 ” として表現することも必要。

参考文献

[1] 藤原 由紀子　五藤 智久　井口 浩人　　コルモゴロフ複雑性に基づく製品・サービスの価値評価
[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)