# 余弦類似度を用いた方言分類木の評価

**An estimation of cosine clustering tree of Japanese dialect**

**谷 研究室　　本荘 智則**

**Tomonori Honjo**

**概要**

Ming Li らにより，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だ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．本研究では圧縮類似度を用いた方言の自動分類の際に作成された木が圧縮類似度をどれほど反映しているかの評価方法を検討する 2007 年度に堀中が定義した平均値や長さで導く木の評価法に加え，類似度判定をベクトルのなす角で導く余弦類似度を用いて，木の評価値を求める実験を行なう．

## 1　はじめに

これまで 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（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]

ところで方言研究では単語単位，アクセント，イントネーションに基づいた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．だが、これまで方言などの類似度判定には専門知識や経験が必要であった．

そこで，谷研究室では近年，専門知識と経験を必要としない圧縮類似度距離を利用した方言の自動分類の研究を行ってきた．今年度も引き続き，方言自動分類の研究を行う．

今年度の実験では，堀中らが音声ソフトをテキスト化したデータに，従来の 2 種類の前処理に加え，新たにライス符号による前処理を追加して全 6 種類の前処理を行うことにした．また，クラスタリングにおいても，これまでの谷研究室の方言自動分類の実験に用いられてきた，NJ 法，UPGMA 法，Quartet Method に加え，新たに連結クラスタリング法を追加して全 4 種類のクラスタリングを行う．更にクラスタリングで得られらた木が圧縮類似度をどれ程反映しているかを検討する．2007 年度谷研究室在籍の堀中が行った ntd(u,v) Quartet Method の S(T) 値，差の平均，2 乗和の総和の平方根の木の評価方法に，新たに余弦類似度を加えて全 4 種類の方法で評価する実験を行う．

第 2 節では圧縮類似度距離について，第 3 節では，NJ 法（近隣結合法）UPGMA 法，Quaret Method の階層型クラスタリングアルゴリズムの解説，第 4 節では木の評価に関しての定義と余弦類似度について，第 5 節では実験概要，実験データについて，第 6 節では実験結果，考察，第 7 節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 $d : X \times X \to R$ 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，$O(logK(xy))$ 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.

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

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 法, Quartet-Method, 連結クラスタリング法 の四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 ([2]).

# 4 木の評価

## 4.1 S(T) 値

Quartet Method における木の評価値である. Quartet Method のコストを計算するには $n^4$ かかってしまい, データが多くなるほど膨大な時間を費やしてしまう. 時間を短縮するために, bestcost(Maximum cost) と worstcost(mimum cost) を計算し, 比較することにする. なので, Quartet Method の cost(以後、C(T) とする) は

必ずこの 2 つの値の間に存在し，その C(T) を評価したのが S(T) 値である．bestcost のとき，つまり良い木の時 S(T)=1 であり，worstcost のとき，つまり悪い木の時 S(T)=0 となる．

### 4.2 木の距離定義

クラスタリングで得られた木が圧縮類似度距離で求めた類似度の距離を，どれ程反映しているかを検証するための評価値を定義する．圧縮類似度距離で求めた類似度の距離を sd(u, v) とし，グラフとしてみた場合の頂点同士の距離（2 頂点間の辺数）を td(u, v) とする．sd(u,v) は圧縮類似度距離で求めた距離表で，td(u,v) はクラスタリングで得られたグラフから定まる距離である．クラスタリングで得られた木が類似度距離をどれ程反映しているかを検証するために sd(u,v) をグラフから定まる距離にする必要がある．そのために td(u, v) を類似度の距離に対応させた距離 ntd(u, v) を定義する．最小値を min(sd)，最大値を max(sd) とする．またグラフとしてみた場合の頂点同士の距離（2 頂点間の辺数）を td(u, v) とし，td(u, v) のうちの最小値を min(td)，最大値を max(td) とする．
td(u, v) を類似度の距離に対応させた距離を ntd(u, v) と定義する．そして変換定数を S とする．

$$\mathrm{S}=\frac{max(sd)-min(sd)}{max(td)-min(td)}$$

$$ntd(u,v)=min(sd)+S(td(u,v)-min(td))$$

### 4.3 木の評価値定義

2007 年度谷研究室在籍の堀中により，ntd(u,v) を用いて，$TV_1$ $TV_2$ の 2 種類の評価値を定義し，3 種類のクラスタリングのどれが圧縮類似度距離に基づいて分類を行っているか客観的な評価値を導く実験を行った． $TV_1$ は差の平均値，$TV_2$ は 2 乗和の総和の平方根である．$TV_1$ は平均を求めているのだが，平均では情報が失われてしまうことがあり，信頼できるデータとは言い切れない．そこで，平均以外で評価できる方法はないかということで，類似度判定に一般的に用いられている，類似度をベクトルのなす角で評価する余弦類似度を新たに加えた．

### 4.4 余弦類似度

$< ntd(u,v), sd(u,v) >$ はベクトル ntd(u,v)，ベクトル sd(u,v) の内積，$||ntd(u,v)||$ は ベクトル ntd(u,v) の長さである．主値は $0 \leqq \cos\theta \leqq 1$ とするのが普通である．ベクトルのなす角が 0 の場合，二つのベクトルは一次従属すなわち方向が同じであり，$\pi/2$ の場合は直交する．この性質から，2 つの零でないベクトルがどれだけ類似しているかの尺度として，ベクトルのなす角 $\theta$ の余弦である次の値をベクトルの類似度とする場合がある．この値は二つのベクトルが一次従属する（もっとも類似している）場合 1，直交する（まったく類似していない）場合 0 になる．

これらから木の評価値を $S(T)$,$TV_1$,$TV_2$,$CV$ と 4 種類定義する．L は末端ノードである．
Quartet Method の評価関数

$$S(T)=\frac{M-C(T)}{M-m}$$

差の平均

$$TV_1=\frac{2}{L(L-1)}\sum_{u,v\,\in\,L}|ntd(u,v)-sd(u,v)|$$

2 乗和の総和の平方根

$$TV_2=\sqrt{\sum_{u,v\,\in\,L}(ntd(u,v)-sd(u,v))^2}$$

余弦類似度

$$CV:\cos\theta=\frac{<ntd(u,v)>\times<sd(u,v)>}{||ntd(u,v)||\times||sd(u,v)||}$$

$TV_1$, $TV_2$ は値が 0 に，$S(T)$、$CV$ は値が 1 に近いほど類似度の距離がより反映されたグラフであることがわかる．

## 5 実験概要

### 5.1 実験データ

方言ももたろう（監修著:杉藤美代子) という日本各地 56 箇所の音声データが入っているソフトに，2007 年度谷研究室在籍の堀中らの手により，音声を手動でテキスト化した文書データがある．そのデータに今年度からは 6 種類の前処理を掛ける．

昨年度までの前処理では，前処理 1 で 50 音順に 1 から対応させて変換し，前処理 2 が使用頻度が高い順に 1 から対応させて変換していく，2 種類であった．今年度

はこの 2 種類の前処理に加えてライス符号を掛けた新たな前処理を加えた．前処理は，ライス符号のパラメータ b の値により，前処理 1(b=8)，前処理 1(b=16)，前処理 2(b=8)，前処理 2(b=16)，前処理 1(符号化無し)，前処理 2(符号化無し) の全 6 種類のデータが存在する．その 6 種類のデータに NJ 法 UPGMA 法 Quartet Method の 3 種類のクラスタリングを掛けたデータを，木の評価値を導くことに使用する．

### 5.2 木の評価

2007 年度谷研究室在籍の堀中により，クラスタリング結果をグラフとして見た場合の頂点同士の距離 (2 頂点間の辺数) と類似度距離の比から導いた ntd(u,v) を用いて，$TV_1$ $TV_2$ の 2 種類の評価値を定義し，3 種類のクラスタリングのどれが縮類似度距離に基づいて分類を行っているか客観的な評価値を導く実験を行った． $TV_1$ は差の平均値，$TV_2$ は 2 乗和の総和の平方根，S(T) 値は Quartet Method の評価関数である． $TV_1$ は平均を求めているのだが，平均では情報が失われてしまうことがあり，信頼できるデータとは言い切れない．そこで，平均以外で評価できる方法はないかということで，類似度判定に一般的に用いられている，類似度をベクトルで評価する余弦類似度を新たに加えてに実験を行う．

全 6 種類の前処理を掛けたデータに 3 種類のクラスタリングを掛けて 4 種類の評価方法で木を評価する．

## 6 実験結果・考察

■符号化無し

| | preprocess1 | preprocess1 | preprocess1 | preprocess1 |
|---|---|---|---|---|
| | $S(T)$ | $TV1$ | $T2$ | $CV$ |
| NJ 法 | 0.427187 | 0.15427 | 1.8345 | 0.351619 |
| UPGMA 法 | 0.427187 | 0.245212 | 16.8321 | 0.0168566 |
| Quartet Method | 0.41834 | 0.256335 | 12.4018 | 0.0199777 |

■符号化 (前処理 1 rice b=8)

| | preprocess1 | preprocess1 | preprocess1 | preprocess1 |
|---|---|---|---|---|
| | $S(T)$ | $TV1$ | $T2$ | $CV$ |
| NJ 法 | 0.576013 | 0.0643147 | 0.727892 | 0.50471 |
| UPGMA 法 | 0.576013 | 0.730638 | 29.7128 | 0.0245355 |
| Quartet Method | 0.472181 | 0.294063 | 13.8375 | 0.0273412 |

■符号化 (前処理 1 rice b=16)

| | preprocess1 | preprocess1 | preprocess1 | preprocess1 |
|---|---|---|---|---|
| | $S(T)$ | $TV1$ | $T2$ | $CV$ |
| NJ 法 | 0.465594 | 0.0573145 | 0.693105 | 0.822158 |
| UPGMA 法 | 0.465594 | 0.282483 | 20.4416 | 0.0392706 |
| Quartet Method | 0.579318 | 0.372329 | 17.5248 | 0.0446572 |

■符号化 (前処理 2 rice b=8)

| | preprocess1 | preprocess1 | preprocess1 | preprocess1 |
|---|---|---|---|---|
| | $S(T)$ | $TV1$ | $T2$ | $CV$ |
| NJ 法 | 0.49459 | 0.0933362 | 1.12188 | 0.725118 |
| UPGMA 法 | 0.49459 | 0.271887 | 19.6051 | 0.0345233 |
| Quartet Method | 0.561546 | 0.360401 | 16.9093 | 0.0385271 |

■符号化 (前処理 2 rice b=16)

| | preprocess1 | preprocess1 | preprocess1 | preprocess1 |
|---|---|---|---|---|
| | $S(T)$ | $TV1$ | $T2$ | $CV$ |
| NJ 法 | 0.439778 | 0.0712431 | 0.808413 | 0.847463 |
| UPGMA 法 | 0.439778 | 0.281534 | 20.4286 | 0.0402256 |
| Quartet Method | 0.61914 | 0.390121 | 18.4503 | 0.0414795 |

余弦類似度でも 2007 年度研究と同じで NJ 法の評価値が一番よかった．それは過去の研究から 3 つののクラスタリングの中では，NJ 法がより類似度距離を反映すると考えられるため．

前処理 2 の評価値が前処理 1 よりよい．それは 50 音順より使用頻度順のデータの方が圧縮後のファイルサイズが小さくなったと考えられるため．

符号化すると評価値がよくなった．それは符号化することにより圧縮前のファイルサイズが小さくなったと考えられるため．

ライス符号の b=16 のとき評価が最もよくなった．それはパラメータを大きくしたことにより更に圧縮前のファイルサイズが小さくなったと考えられるため．

## 7 今後の課題

他のクラスタリング方法でも木の評価値を出す．

例：連結クラスタリング法

他の圧縮方法でも木の評価を行う．

## 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] 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