情報処理学会論文誌 Vol. 49 No. 12 4016–4024 (Dec. 2008)

正規圧縮距離を用いた和文小説の著者別分類と圧縮プログラムの妥当性

石原正道[†1] 佐藤静香[†1]

本研究では和文小説の著者別分類において正規圧縮距離（NCD）を用いた距離行列法が有効であるか調べた．距離行列のみでは文章間の関係が確定しないため，非加重平均結合法（UPGMA 法）・近隣結合法（NJ 法）・古典的多次元尺度構成法（MDS 法）により文章間の関係を推定することを試みた．これらの方法が有用であるか判断するために，ある和文から順次改編を行い作成した文章群に対し NCD を適用して距離行列を作成し，これらに UPGMA 法・NJ 法・MDS 法を適用して文章間の関係を推定した．また和文小説への適用可能性を調べるために，圧縮プログラム bzip2 および gzip を利用し，小説のデータから NCD による距離行列を求めた．本研究では距離行列の対角要素を利用した圧縮プログラムの選択方法について考察した．また距離行列に NJ 法および MDS 法を適用し，著者別の分類が可能であるか調べた．以上により，本研究では NCD で用いられる圧縮プログラムの選択基準として，距離行列の対角要素の値の分布において，平均値が小さいことおよび分布の幅が狭いことを条件として提案した．また NJ および MDS を用いることで，NCD による距離行列法が和文小説の著者別分類に有用であることを示した．

Classification of Japanese Novels according to Authors by Normalized Compression Distance and Validity of Compression Programs

Masamichi Ishihara[†1] and Shizuka Sato[†1]

We studied the applicability of distance method by using Normalized Compress Distance (NCD) for the classification of Japanese novels according to authors. The methods to extract the relations among the texts from a distance matrix were studied, because the distances do not determine the relations completely. Therefore we constructed a series of texts which is made from a certain text, and studied, by drawing the diagrams, the effectiveness of some methods which are unweighted pair group method using arithmetic averages (UPGMA), neighboring joining method (NJ) and classical multidimensional scaling method (MDS). In addition, we investigated the applicability of the distance given by NCD for the classification of Japanese novels. Therefore we construct the distance matrix of Japanese novels by NCD with bzip2 and gzip. We propose criteria to select a program for compression. The criteria are 1) the width of the distribution of diagonal elements is narrow and 2) the average value of the diagonal elements is small. We also found that Japanese novels are classified by authors by using normalized compression distance with the help of the diagrams depicted by NJ and MDS.

1. はじめに

近年，和文の分類において，助詞や句点の出現頻度などの統計を用いた分類方法が導入され，多くの成果があげられている[1)–5)]．統計手法に類する客観的な手法に基づく和文の分類方法の 1 つに，文章間の距離（非類似度）を定義し，定義された距離を用いて分類する距離行列法がある．文書の分類に利用される距離としてはコサイン係数を用いた距離などがある．また近年用いられるようになった距離としては正規圧縮距離[6)]（NCD）がある．NCD はコルモゴロフ複雑量を基礎としており，遺伝子における解析[7)]，ゲノムのデータを利用した系統樹作成[8)]，欧米の文学作品の分類[9)] や音楽の分類[9),10)] などに適用され，おおむね良好な結果が得られている．NCD は圧縮後のデータ量から距離を計算するため，対象に対する予備知識を必要とせず，ベクトル化の必要もないという利点を持っている．

これまでの成功例から NCD を用いた距離行列法は和文の分類にも適用可能であると思われる．しかし日本語には数多くの文字があり，欧文に適用できても和文にも適用できるとは限らない．また NCD から得られる文章間の距離から，複数の文章間の関係をどのように推定すべきかも明白でない．そこで本研究では，明白な順序関係を持つ文章を用いて，NCD を用いて得られる距離行列が文章を分類をするうえで有用であるか調べた．またこの過程において距離行列から文書間の関係を推定するための方法が必要である．そこで本論文ではこれらの方法に非加重平均結合法[11)]（UPGMA 法）・近隣結合法[11)]（NJ 法）・古典的多次元尺度構成法（MDS 法）を選び，これらの方法のうち，どの方法が文章間の関係の推定に有効な方法であるか検討した．さらに，より一般的な文章においても NCD を用いた分類が有用であるかを判断するため，著者が既知である文章データ（小説）に対し NCD を用いた距

†1 郡山女子大学人間生活学科
Department of Human Life Studies, Koriyama Women's University

離行列法を適用し，分類できるか調べた．

2 章では，まず NCD の定義および NCD より得られる距離行列について触れた．また本章において，距離行列から文章間の関係を推定する方法の選定を行った．3 章では NCD を用いて得られる距離行列と 2 章で選定された推定方法を用い，著者が既知の小説に対して，NCD により著者ごとの分類が可能であるかについて調べた．4 章では本研究によって得られた結論を記した．

2. 文章間の関係の推定

2.1 正規圧縮距離と距離行列

正規圧縮距離（NCD）は正規情報距離に基礎をおいている．ある文章 x のコルモゴロフ複雑量 $K(x)$ を用いて，正規情報距離は定義される[6),8)]．入力のない，x を出力するプログラムを考えると，このプログラムにより x の情報が表現されていると考えられる．このため，これらのプログラムのうちの最小サイズのプログラムには，x の情報が集約されているといえる．コルモゴロフ複雑量 $K(x)$ はこの最小のプログラムサイズとして定義される．実用上は $K(x)$ を可逆圧縮プログラム C により圧縮した情報量 $C(x)$ で近似する．2 つの文章 x，y に対し，この $C(x)$ を用いて NCD は

$$\mathrm{NCD}(x,y) := \frac{\max\left[C(xy)-C(x), C(yx)-C(y)\right]}{\max\left[C(x), C(y)\right]}$$

と定義される[12)]．定義式中の xy の情報には x の情報が含まれている．そこで xy の情報量 $C(xy)$ から x の情報量 $C(x)$ を引くと，y の情報量のうちで x によって記述できない情報量を抽出したことになる．また $\max\left[C(x), C(y)\right]$ による除算により，x および y そのものの情報量による影響を排している．理想的に圧縮ができた場合には，$C(xx) \sim C(x)$ であると考えられるから $\mathrm{NCD}(x,x) \sim 0$ となる．また y の情報に x の情報がまったく含まれていない場合，すなわち $C(xy) - C(x) \sim C(y)$ の場合には $\mathrm{NCD}(x,y) \sim 1$ となることも分かる．

NCD を距離として利用するためには $\mathrm{NCD}(x,x)$ が零となることが望まれるが，実際の文章において零となることは期待できない．したがって，$\mathrm{NCD}(x,x)$ が $\mathrm{NCD}(x,y)$（$x \neq y$）より十分小さいならば満足しなければならない．圧縮プログラム C としては多くの候補があるが，本研究では圧縮率が高いとされる bzip2 を利用した．本論文内で特に断らなければ bzip2 による結果である．また比較として gzip も用いている．以下の結果において gzip を用いている場合は，圧縮プログラムを明示する．

文章 x_i と x_j の NCD による距離を $d_{ij} := \mathrm{NCD}(x_i, x_j)$ とする．距離行列 D の ij 成分は d_{ij} により与えられる．距離行列の対角要素は零となることが望ましいが，一般に $d_{ii} \neq 0$ となっている．多くの場合，相異なる文書間の距離 d_{ij}（$i \neq j$）に比較して d_{ii} は十分小さくなっており，本論文では条件 $d_{ii} = 0$ は近似的に満たされているものとして扱う．明らかに $d_{ij} = d_{ji}$ であるため，行列 D は実エルミート行列となる．またコンピュータ上の和文では，複数の文字コードが利用されている．このため文章ごとに文字コードが異なると，d_{ij} を一意に確定できなくなりうる．そこで本論文で扱う文章では文字コードを SJIS に統一して扱った．以下ではこの距離 d_{ij} より距離行列 D を作成し，文章間の関係を推定した．

2.2 文章の相互関係の推定方法

距離行列 D が与えられたとき，文章間の関係を推定する方法として様々なクラスタ法が広く用いられている．本節では，どのような方法を用いると文章間の関係が明らかになりやすいか調べた結果を示す．そこでまず 1 編の和文を用意し，文章の一部を任意に選択して変更し，一連の文章群を作成した．2 つの連続したファイルにおける差は全角文字で数文字程度の違いである．当然ながら，これらの差の大きさは差分ごとに異なっている．起点となる文章を 1.txt とし，1.txt を元に改変した文章を 2.txt，2.txt を元に改変した文章を 3.txt とファイル名をつけていき，最終的に 11 ファイルを順次作成していった．**表 1** に使用した文章のデータサイズを示す．また変更例を**表 2** に示す．この方法により作成した 11 ファイルを用い，NCD によって得られた距離行列から元の文章の列を推定するための方法として UPGMA 法，NJ 法，MDS 法のいずれが良いか検討した．UPGMA 法ではクラスタ間の距離をクラスタ内の要素間の平均距離により決定し，距離の近いクラスタを併合していくことでクラスタリングを行う．NJ 法では，2 つのクラスタ間の距離に他のクラスタま

表 1　文章のデータサイズ．ファイルは改変した順序に従って番号づけされている
Table 1　Data sizes of the writings. The files are numbered by the order of the modifications.

ファイル番号	1	2	3	4	5	6	7	8	9	10	11
ファイルサイズ（バイト）	826	804	804	748	746	746	748	748	748	750	744

表 2　文章の変更例．本表では 1.txt と 2.txt および，2.txt と 3.txt の間の変更例を示した
Table 2　Examples of the modifications of Japanese texts. In this table, we display the difference between 1.txt and 2.txt, and that between 2.txt and 3.txt.

	変更前の文章	変更後の文章
1.txt→2.txt	秋元様	[削除]
	貴社ますますご清栄のこととお喜び申し上げます．	貴社益々ご清栄のこととお喜び申し上げます．
2.txt→3.txt	拝啓	謹啓
	敬具	敬白

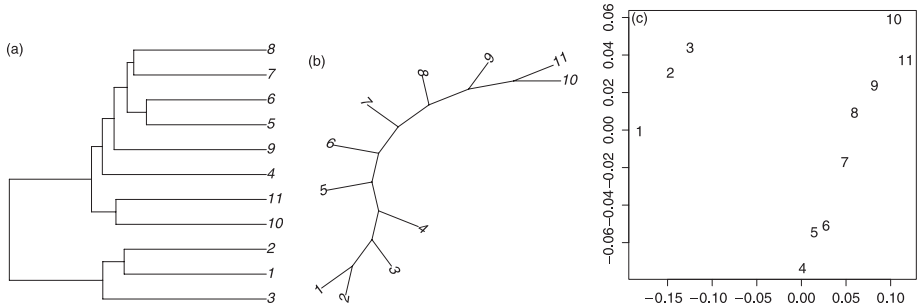


図 1　NCD を用いた，11 編に対する文章の関係の推定．順序関係の明白な文章に対し，文書間の距離を NCD によって決め，(a) 非加重平均結合法，(b) 近隣結合法，(c) 多次元尺度構成法により関係を推定した

Fig. 1　Estimation of the relations of 11 texts by NCD. The relations among the ordered texts are estimated by NCD with (a) UPGMA, (b) NJ and (c) MDS.

での距離を考慮して距離を補正し，この補正された距離を用いてクラスタリングを行う．一方 MDS 法では可能な限り要素間の距離を再現するように多次元空間内に要素を配置する．これらの解析には統計解析ソフト R バージョン 2.6.0 および ape パッケージ（ver. 2.0-1）と stats パッケージ（ver. 2.6.0）を使用した．本論文の計算では距離 d_{ii} が非零となる問題があるが，実際に距離行列を求めると対角要素のほとんどは非対角要素と比べて十分小さくなっていた．そこで本研究ではこの対角要素は文章間の関係を求める際には利用していない．すなわち非対角要素 d_{ij}（$i > j$）のみを用いている[⋆1]．

図 1 (a)，(b)，(c) は，圧縮プログラムを bzip2 とし，(a) UPGMA 法，(b) NJ 法，(c) MDS 法により，先の 11 ファイルの関係を描画したものである．したがって 1.txt から 11.txt まで順に並べば再現性が良いことになる．図 1 (a) では明らかに番号順には結合されておらず，再現性が十分であるとはいえない．これに対し図 1 (b) では 1.txt から 11.txt まで番号順に結合されており，きわめて再現性が高い．図 1 (c) ではおおむね番号の近い文章が近い位置にある．3.txt と 4.txt の間に飛びがあるが，表 1 に示されているように 3.txt と 4.txt の間では差は比較的大きく，MDS 法による結果はこの事実を反映しているものと思われる．また圧縮プログラム C を gzip とした場合でも，UPGMA 法・NJ 法・MDS 法でほぼ同様の結果が得られた．

この文章群に加え，起点となるファイルのサイズが 429 バイトから 4,293 バイトである

⋆1 3 章で示すように，対角要素は圧縮プログラム C が妥当であるか判断するために利用できる．

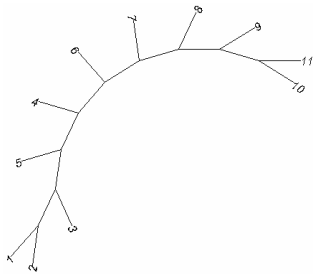


図 2　NJ 法において，一部の順序が逆に接続された図の一例

Fig. 2　An example of the graph with NJ. In this graph, the order is partially interchanged.

9 つの文章群に対し，同様の分析[⋆2]を行った（連続する文章間の差異が十文字以上のものも含む）．UPGMA 法および MDS 法では先の結果と同様の結果となった．ただし MDS 法では平面上にプロットするため，順番の分かりにくい例がいくつかみられた．また NJ 法ではほとんどの文章群で番号順に結合された．一部の文章群で番号順が逆になった箇所もあったがごくわずかにすぎない．一例を**図 2** に示す．

以上から，NJ 法による結果を主として参照し，MDS 法により得られる知見を補足すると良好な結果が得られと予想される．

3. 小説の著者別分類

3.1 小説のデータ

本章では小説データの著者別分類を試みる．小説を著者ごとに分類する場合，NJ 法を用いて推定された図では，つながれた枝を切断し，どこまでが同一の著者の作品であるか判断をすることは難しい．そこで NCD を用いて距離を算出し，NJ 法により得られる結果と MDS 法により得られる結果を併用することで文章間の関係を推定する．

小説のデータはインターネット上の青空文庫[13] に登録されているデータを用いた．このデータにはファイル内の先頭部および最後部に電子化した際に加えた情報がある．この部分は本来の小説データでないため，取り除いて距離の算出を行った．また本文中のルビは取り除いていない．利用した小説の著者は森鴎外・太宰治・芥川龍之介・夏目漱石の 4 人であり，1 人あたり 25 編の小説を用い，全 100 編の小説を利用した．データはファイルサイ

⋆2 パッケージが更新されたため，ここでは ape パッケージ ver. 2.2 を使用した．

表 3　著者別の分類に使用した小説．ファイルサイズは 20 KB 以上である．森鴎外の作品には 1 番から 25 番，太宰治の作品には 26 番から 50 番，芥川龍之介の作品には 51 番から 75 番，夏目漱石の作品には 76 番から 100 番が表の並びの順に割り当てられている

Table 3　List of novels for the classification by authors' names. The sizes of these novels are greater than or equal to 20 KB. These novels are numbered in order of the list. The novels by Ougai Mori are numbered from 1 to 25, those by Osamu Dazai from 26 to 50, those by Ryunosuke Akutagawa from 51 to 75 and those by Souseki Natsume from 76 to 100.

著者	作品
森 鴎外	高瀬舟，沈黙の塔，舞姫，雁，寒山拾得，牛鍋，魚玄機，栗山大膳，里芋の芽と不動の目，山椒大夫，食堂，心中，青年，独身，鶏，花子，百物語，二人の友，空車，妄想，安井夫人，余興，追儺，そめちがえ，椙原品
太宰 治	五所川原，列車，愛と美について，お伽草子，人間失格，走れメロス，母，眉山，ろまん燈籠，逆行，青森，朝，あさましきもの，兄たち，或る忠告，一燈，陰火，嘘，花燭，風の便り，喝采，家庭の幸福，狂言の神，グッド・バイ，乞食学生
芥川 龍之介	秋，アグニの神，ある頃の自分の事，或阿呆の一生，或敵打の話，糸女覚え書，糸と笛，馬の脚，芋粥，老いたる素，お律と子等と，開化の良人，開化の殺人，影，河童，神々の微笑，枯野抄，木曽義仲論，疑惑，戯作三昧，玄鶴山房，好色，湖南の扇，西方の人，邪宗門
夏目 漱石	永日小品，思い出すこと，カイ露行，硝子戸の中，草枕，虞美人草，現代日本の開化，行人，抗夫，こころ，琴のそら音，三四郎，自転車日記，趣味の遺伝，創作家の態度，それから，手紙，点頭録，道楽と職業，中味と形式，野分，彼岸過迄，文芸と道徳，文鳥，道草

ズが 20 KB 以上のものを使用した．この理由は非常に小さいファイルサイズの文章データでは文章間の類似性よりも圧縮プログラム C の影響が現れやすいと考えられるためである．またファイルサイズが大きいほど，著者の情報を多く含むと考えられるので，ファイルサイズが 100 KB 以上の 18 編の小説の分類も試みた．**表 3** に利用した 20 KB 以上の 100 編の小説を示す．また**表 4** に利用した 100 KB 以上の 18 編の小説を示す．

3.2　圧縮プログラムの妥当性

圧縮プログラム C の妥当性を調べるため 100 編の小説から距離 d_{ij} を成分とする行列 $[d_{ij}]$ を作成し，対角要素の値 d_{ii} からヒストグラムを作成した．**図 3** (a)，(b)，(c) はそれぞれ圧縮プログラムを bzip2，gzip，gzip -9 とした場合に対するヒストグラムである．横軸に対角要素の値をとり，縦軸に頻度をとった．また階級数を 40，階級幅を 0.025 ととった．図 3 (a) では d_{ii} の値が 0.2 近傍に集中しているのに対し，図 3 (b)，(c) では 0 と 1 の付近に二極化している$^{\star 1}$．対角要素の値 d_{ii} が文章により大きく変動するということは，文章に対する依存性が大きいことを意味する．安定した結果を得るにはこの変動が小さい方が良く，図 3 (b)，(c) から本節での分類において gzip は適当でないことが分かる．これに対し bzip2 では文章の違いによる変動は比較的小さい．よって本研究で使用した文章を NCD で分類するためには gzip より bzip2 を用いた方が良いといえる．図 3 にあるように，対角要素が零とならないことは距離行列の性質としてみると望ましくはないが，その一方，対角要素の値の分布を圧縮プログラムを選択するために用いることができる．

$\star$1 対角要素 d_{ii} が 1 に近い値をとる理由の 1 つとして，具体的な圧縮プログラムにおいては有限の Window Size を用いて圧縮が行われる点が指摘されている[14]．対角要素 d_{ii} の計算において Window Size を超えたデータを圧縮する場合に，d_{ii} は 1 に近い値をとりうることが示されている．

表 4　著者別の分類に使用した小説．ファイルサイズは 100 KB 以上である．1 番から 5 番は森鴎外，6 番から 10 番は太宰治，11 番から 13 番は芥川龍之介，14 番から 18 番は夏目漱石の作品である

Table 4　List of novels for the classification by authors' names. The sizes of these novels are greater than or equal to 100 KB. The novels by Ougai Mori are numbered from 1 to 5, those by Osamu Dazai from 6 to 10, those by Ryunosuke Akutagawa from 11 to 13 and those by Souseki Natsume from 14 to 18.

番号	作品	番号	作品
1	雁	10	右大臣実朝
2	キタ・セクスアリス	11	路上
3	伊沢蘭軒	12	鑑盗
4	大塩平八郎	13	素戔嗚尊
5	青年	14	我輩は猫である
6	お伽草子	15	それから
7	人間失格	16	こころ
8	斜陽	17	彼岸過迄
9	新ハムレット	18	行人

また NCD による距離が妥当であるためには，行列 $[d_{ij}]$ の対角要素が非対角要素より十分小さい必要があった．**図 4** は圧縮アルゴリズムに bzip2 を使用して得られた非対角要素の値から作成したヒストグラムである．図 4 から非対角要素の値は 1 よりやや小さい値の近傍に集中していることが分かる．このことから対角要素は非対角要素に比べ十分小さいことが分かる．よって対角要素 d_{ii} を零として扱うことが可能であると考えられるから，$[d_{ij}]$ を距離行列として扱うことは妥当である．

3.3　著者別分類における推定方法

本節では NCD を用いて距離行列から NJ 法および MDS 法により小説間の関係を描画した．小説の著者は未知であるものとして小説を番号で表すことにする．またファイルサイズが大きいほど著者の特徴差が現れやすいと考えられる．そこでまず，ファイルサイズが 100 KB 以上の 18 編の小説を分類した結果を**図 5** に示す．図 5 (a)，(b) から文章 $(6, 7, 8, 9, 10)$ は組であると判別できる．同様に $(1, 2, 5)$，$(16, 17, 18)$ はそれぞれ組になっているといえるだ

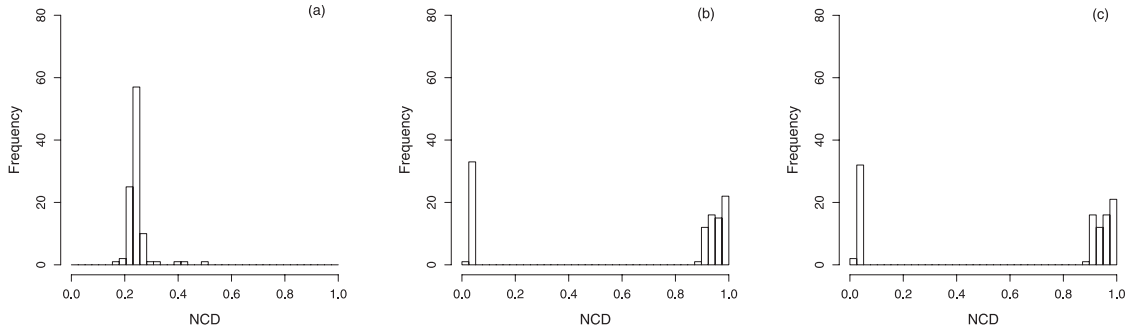


図 3　行列 $[d_{ij}]$ の対角要素の値によるヒストグラム．階級数を 40，階級幅を 0.025 とした．対角要素数は 100 であり，圧縮プログラムは (a) bzip2，(b) gzip，(c) gzip -9 である
Fig. 3　Histograms of the diagonal elements of the matrix $[d_{ij}]$. The number of bins is 40 and the widths of the bins are 0.025 in each figure. The number of the diagonal elements is 100. The compressors for these figures are (a) bzip2, (b) gzip and (c) gzip -9, respectively.

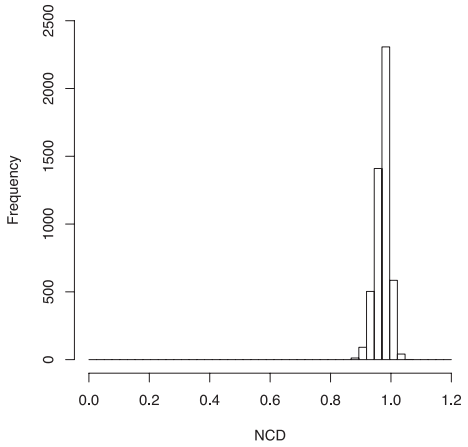


図 4　100 編の小説に対する行列 $[d_{ij}]$ の非対角要素の値によるヒストグラム．階級数を 48，階級幅を 0.025 とした．ヒストグラムを作成するために用いた非対角要素数は 4950 である
Fig. 4　Histograms of the off-diagonal elements of the matrix $[d_{ij}]$ for 100 novels. The number of bins is 48 and the widths of the bins are 0.025. The number of the off-diagonal elements which are used to draw this histogram is 4950.

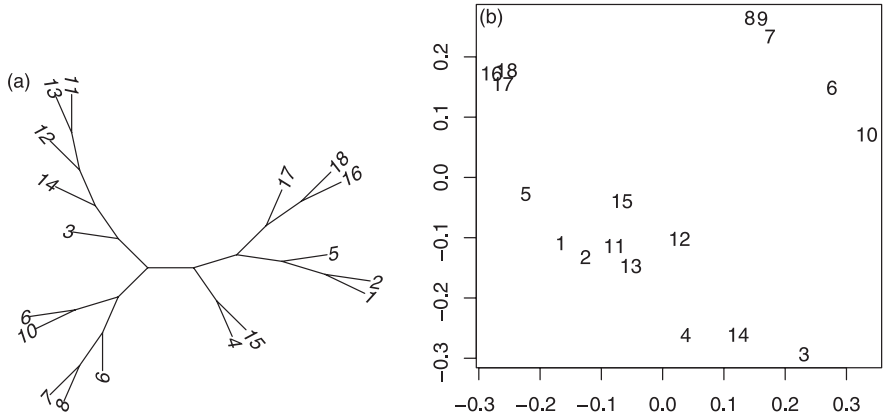


図 5　100 KB 以上の小説 18 編の分類結果．分類には (a) NJ 法および (b) MDS 法を用いた
Fig. 5　Classification of 18 Japanese novels by (a) NJ and (b) MDS. The sizes are larger than or equal to 100 KB.

ろう．図 5 (a) から $(11, 12, 13, 14, 3)$ の組となっているようにも見えるが，図 5 (b) を考慮すると，$(11, 12, 13)$ と 3，14 は分けるべきであろう．したがって $(1, 2, 5)$，$(6, 7, 8, 9, 10)$，$(11, 12, 13)$，$(16, 17, 18)$ がそれぞれ 1 人の著者を表しているとみてよい．実際，各々の組は 1 人の著者の作品群に属している（表 4）．残った文章 3，4，14，15 の分類は困難である．たとえば文章 15 は図 5 (a) から $(16, 17, 18)$ に属するとするのか，図 5 (b) から $(1, 2, 5)$ あるいは $(11, 12, 13)$ に属するとするのか判断することは難しい．これらの文章では，bzip2 により計算した NCD を用いた NJ 法あるいは MDS 法のみによる分類では誤りかねないことが分かる．原因の 1 つとして，同一の文章間の NCD の値の大きさがあげられる．この

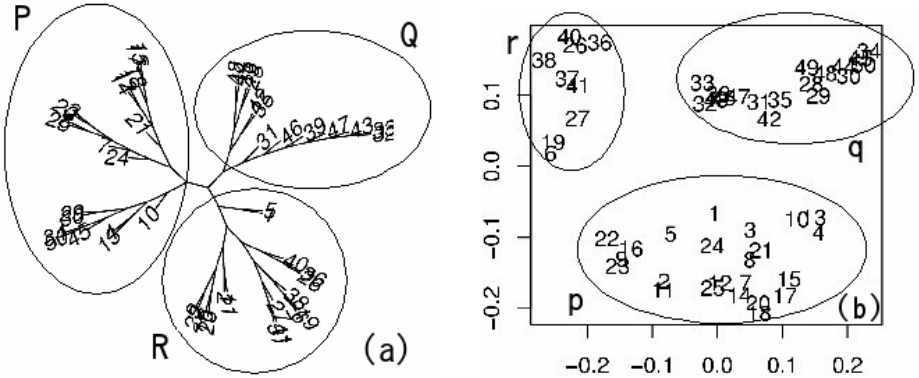


図 6　森鴎外および太宰治の 20 KB 以上の小説の分類．(a) は NJ 法，(b) は MDS 法により描画した
Fig. 6　Classification of Japanese novels by (a) NJ and (b) MDS. The authors are Ougai Mori and Osamu Dazai. The sizes are larger than or equal to 20 KB.

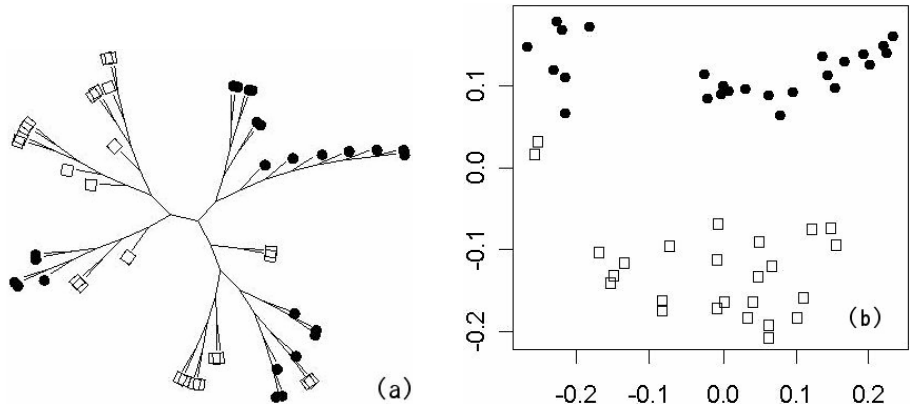


図 7　森鴎外と太宰治の小説に対し (a) NJ 法，(b) MDS 法により得た小説間の関係図．四角の記号は森鴎外，丸の記号は太宰治の小説を表している
Fig. 7　Classification of Japanese novels by (a) NJ and (b) MDS. Squares and black circles represent the novels by Ougai Mori and those by Osamu Dazai, respectively.

値は文章 3 では約 0.99，文章 14 では約 0.89 となっており，このために適切な分類ができていない可能性がある．以上から分類しきれない文章はあるものの，他の知識をいっさい利用せず，NCD による距離を用いることで和文において著者別の分類が可能であることが分かる．

次に 20 KB 以上の小説に対し，NCD による距離行列から (a) NJ 法および (b) MDS 法により作品間の関係を描画した図を示す．4 著者の文章を同時に扱うと数が多すぎ，関係が判然としなくなる．そこで 2 著者を組とし，各々について文章間の関係を描画した．著者の組合せは 6 通りであるが，ここではそのうちの一部を図として示す．**図 6** では森鴎外と太宰治の各 25 編計 50 編に対し，作品間の関係を描画した．(a) は NJ 法，(b) は MDS 法による結果である．以下では枝や領域を表すために [···] という記法を用いる．図 6 (a) では左部の枝 [P]，右上の枝 [Q]，右下の枝 [R] に分かれている．図 6 (b) では下部 [p]，右上部 [q]，左上部 [r] の 3 つの部分に分かれている．図から [P] は [p]，[Q] は [q]，[R] は [r] にほぼ対応していることが分かる．2 人の著者であることを既知とし，それぞれの著者による小説の数がおおむね等しい場合を考える．NJ 法ではどこで枝を切断して著者別に分類すればよいか分かりにくい．そこで MDS 法の結果を援用する．[p][q][r] のうち，[p] と [q] に属する小説の数が多い．著作数がおおむね等しいのであるから [p] と [q] は異なる著者によるものであると判断できる．さらに著者は 2 人であるから，[r] を [p] または [q] とまとめる必要がある．[p][q][r] に属する小説の座標値から重心を求めると，それぞれ $(0.065, -0.1345)$，$(0.1064, 0.1144)$，$(-0.2295, 0.1150)$ である．これから [q][r] の距離の方が [p][r] 間よりわずかに近い．したがって，もし重心間の距離の近いクラスタをまとめるとするならば，[Q][R] および [q][r] をまとめることになる[⋆1]．これにより分類は [P][Q,R] および [p][q,r] となり，おおむね小説を著者別に分類できることが分かる．ここまでは著者が不明であることを前提として小説を番号で表していたが，もし著者が既知であり，著者間の関係を求めるという視点であれば，著者が同一である小説をまとめて同じ記号を使う方が分かりやすい．そこで図 6 から著者別に記号化し作成した図を**図 7** に示す．図 7 から小説は著者別に分離されていることが分かる．

前例では分類するうえで MDS 法が有効であったが，多くの場合において図 6 (b) ほど明確に小説は分離されない．そこで他の例として森鴎外と芥川龍之介の場合を**図 8** に示す．図 8 (b) から MDS 法のみを用いて分類をすることは難しいことが分かる．図 8 (a) ではいくつかのグループとなるであろう枝が複数ある．もし小説が 2 著者によるものであるとす

⋆1 重心間の距離の関係は [p][q] 間距離 < [q][r] 間距離 < [p][r] 間距離となっている．それゆえ 2 著者の小説数がおおむね等しいという条件がなければ，重心間の距離の近いクラスタをまとめるという方法では [p] と [q] をまとめることになる．

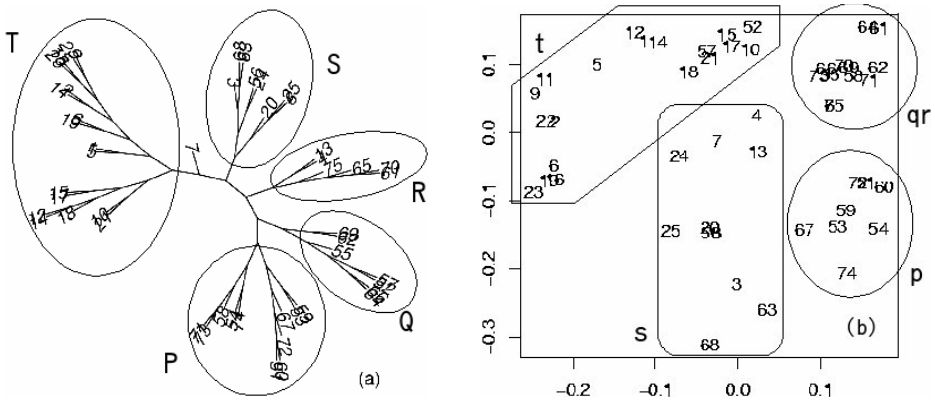


図 8　森鴎外および芥川龍之介の 20 KB 以上の小説の分類．(a) は NJ 法，(b) は MDS 法により描画した
Fig. 8　Classification of Japanese novels by (a) NJ and (b) MDS. The authors are Ougai Mori and Ryunosuke Akutagawa. The sizes are larger than or equal to 20 KB.

れば，どこかで切断をすることになるが，NJ 法の結果だけでこれを判断することは困難である．そこである程度グループ化されていることを考慮し，MDS 法による結果を取り入れて切断点を決定する．ここで図 8 (a) に示したグループを [P][Q][R][S][T] と名付ける．同様に図 8 (b) に示したグループを [p][qr][s][t] と名付ける．

図 8 (a) で [P][Q][R] は近く，また図 8 (b) でも [p][qr] は近いため，これらのグループに属する小説の著者は同一であると考えられる．[Q][R][S] は図 8 (a) では近いが，図 8 (b) で [qr] と [s] はやや離れており，ここが枝を切断する候補点の 1 つとなる．同様に [s][t] もやや離れており，[S][T] の間も切断する候補点であろう．いずれか一方を選ぶならば，文章 7 が [S][T] 間にあること，文章 7 は [s][t] のいずれにも属するようにも見えることから，[S][T] を組とし切断は [R][S] の間ということになろう．あるいは前例と同様に，2 著者の作品数がおおむね等しいならば，作品数から切断は [R][S] の間であると判断できる．どちらの場合でも [P,Q,R][S,T] および [p,qr][s,t] と分離される．これらの方法により，おおむね小説を著者別に分類できることが分かる．前例と同様に，著者別に記号化し作成した図を**図 9** に示す．この図からも，おおむね小説は著者別に分離されていることが分かる．

以上ではファイルサイズが 20 KB 程度以上かつ比較的短い文章に対し，NCD を用いた距離行列法を適用した．本章の結果は，本方法により小説を著者別に分類できることを示している．

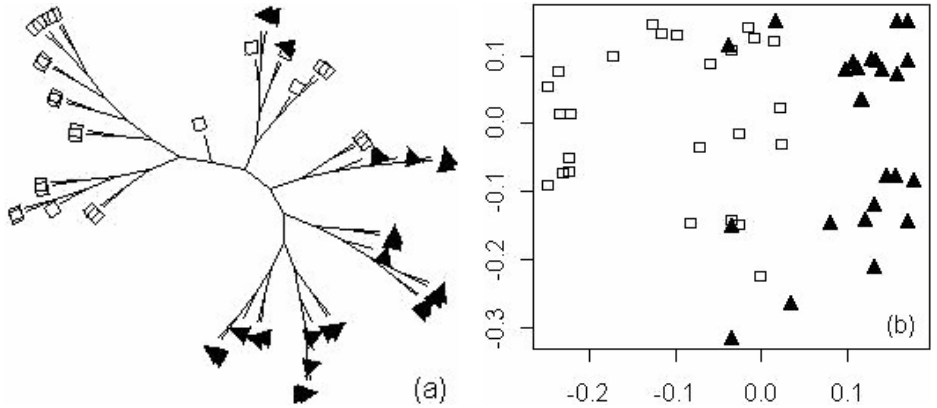


図 9　森鴎外と芥川龍之介の小説に対し (a) NJ 法，(b) MDS 法により得た小説間の関係図．四角の記号は森鴎外，三角の記号は芥川龍之介の小説を表している
Fig. 9　Classification of Japanese novels by (a) NJ and (b) MDS. The authors are Ougai Mori and Ryunosuke Akutagawa. Squares and triangles represent the novels by Ougai Mori, and those by Ryunosuke Akutagawa, respectively.

4. 結　　論

本研究では正規圧縮距離（NCD）を用いた距離行列法により，和文の分類が可能であるか調べた．距離行列法には非加重平均結合法（UPGMA 法）・近隣結合法（NJ 法）・古典的多次元尺度構成法（MDS 法）を用いた．これらの手法により文章間の関係を得ることが可能か調べた．

まず，わずかずつ順序だって改変されている文章群に対しては，NJ 法および MDS 法により改変順序を十分良く再構成できることが分かった．

次に 4 著者の小説を用い，これらの小説を著者別に分類できるか調べた．この著者別の分類に先立ち，圧縮プログラム C の選定方法について考察した．コルモゴロフ複雑量を近似するという観点からは高圧縮である圧縮プログラムを使う方が良い．一方で NCD による同一文章間距離の文章依存性は小さくなければならない．また 3.2 節で示したように圧縮プログラムによっては文章に対する依存性が大きい．そこで「距離行列の対角要素の値の分布において，平均値および分布の広がりが小さくなる圧縮プログラムを選択する」という基準を提案したい．

著者別の分類においては，まず著者が未知である場合に著者別の分類が可能か調べた．著者が未知であると仮定して小説を番号で表し，NCD による距離行列から NJ 法と MDS 法を併用し小説間の関係を推定した．この結果からサイズの大きい小説（100 KB 以上）に対しては著者別の分類が可能であることが判明した．またサイズがやや小さい（20 KB 以上）小説においては，判断が難しい文章があるものの，おおむね文章の分類が可能であることが判明した．次に著者が既知である場合に著者別の分類が可能か考察した．著者ごとに記号を付し，著者が未知として作成した文章間の関係図において，小説の番号を小説の著者の記号に置き換えた．この結果，サイズがやや小さい小説においてさえ，おおむね著者別に分類できることが分かった．1 編の，著者が不明である小説を誰の著作物か判断するには，著者が既知である小説を分類した図において，著者が不明である小説の位置を決めればよい．この方法によりその小説が誰の著作物か判断できる．これらの結果から，著者が未知である複数の小説に対しては完全に分類することは困難な場合がありうるが，多くの場合は NCD を用いた距離行列法は和文の分類において有効であることが明らかになった．とりわけ，著者が不明である 1 編の作品を，誰の著作物であるか判断する際に有効であると考えられる．

本論文における結果から，NCD を用いた分類法は十分良く機能するといえる．しかし前章における分類結果では明確に著者ごとに分類できていない作品もあり，従来の分類手法に比べ，NCD を用いた分類の方が必ずしも有利であるとはいえない[5]．それにもかかわらず NCD を用いる利点としては，他の方法における問題点が NCD を用いる方法では存在しないことがあげられる．たとえば品詞などの分布の差異により区別をする場合，差を検出できるか否かは単語などのサンプル数に依存してしまう．短い文章に対してはサンプル数が少なくなり差の検出が難しくなる一方で，サンプル数が非常に多い場合にはごくわずかな差でも分布が異なると判定されてしまう[15]．また分類の基準（分布を用いた分類における品詞の選択など）の取り方によっては分類に必要な情報が欠落しうる．これらの問題は本論文で用いた方法では生じえず，NCD による分類法は相補的な役割を果たしうる．

NCD は文章を圧縮しさえすれば求められるため，未知の単語が含まれていたり文法構造がはっきりしていない文章の分類に適用できる．現代小説や Web 上の文章では新しい言葉が用いられたり文法が守られていなかったりすることがあるが，これらの文章群も NCD を用いて分類を行うことが可能である．これらの文章群を NCD を用いて明確に分類できるかを調べることは今後の課題であろう．

本論文で示したことは以下の点にまとめられる．1) ある和文から順次変更が加えられていった文章群に対しては，NCD を距離とし NJ 法を用いることで文章の改変順序をほぼ特定できることを示した．2) これまで欧文において NCD による分類が有効であることが示されていた．これに対し，本論文では漢字など非常に多くの文字を有しかつ文法の異なる和文小説の分類において，NCD を距離とした NJ 法および MDS 法を用いた距離行列法が有効であることを示した．この結果は和文の分類に対し，NCD が有効であることを強く示唆するものである．3) NCD による分類を行ううえで圧縮プログラムが妥当であるかは重要な点である．これまで圧縮プログラムの検討は，いくつかのデータを用いてわずかになされているのみであった[14]．特に多くのデータに対する距離行列を用いた評価はこれまでなされていない．本論文では距離行列の対角要素の値の分布，特に平均値と分布の幅により，圧縮プログラムの妥当性を検証すべきであることを具体的な事例で示した．

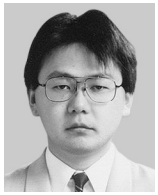
本論文では和文のみを対象としたが，得られた結果は欧文以外の他言語においても NCD を用いた分類が可能であることを示唆するものである．本研究が和文の分類における一助となることを期待したい．

参 考 文 献

1) 村上征勝，今西祐一郎：源氏物語の助動詞の計量分析，情報処理学会論文誌，Vol.40, No.3, pp.774–782 (1999).
2) 村上征勝：文化を計る—文化計量学序説，朝倉書店 (2002).
3) 村上征勝：シェークスピアは誰ですか？ 計量文献学の世界，文藝春秋 (2004).
4) 金　明哲：自然言語における統計手法を用いた情報処理，統計数理，Vol.48, No.2, pp.271–287 (2000).
5) 金　明哲：自己組織マップと助詞分布を用いた書き手の識別と特徴分析，日本行動計量学会大会発表論文抄録集，Vol.30, pp.194–197 (2002).
6) Li, M., Chen, X., Li, X., Ma, B. and Vitányi, P.: The Similarity Metric, *IEEE Trans. Inf. Theory*, Vol.50, No.12, pp.3250–3264 (2004).
7) Nykter, M., Yli-Harja, O. and Shmulevich, I.: Normalized compression distance for gene expression analysis, *Workshop on Genomic Signal Processing and Statistics* (*GENSIPS*) (2005).
8) Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P. and Zhan, H.: An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *BIOINFORMATICS*, Vol.17, No.2, pp.149–154 (2001).
9) Cilibrasi, R. and Vitányi, P.: Similarity of Objects and the Meaning of Words. arXiv:cs/0602065
10) Cilibrasi, R., Vitányi, P. and de Wolf, R.: Algorithmic Clustering of Music. arXiv:cs/0303025v1
11) 阿久津達也：バイオインフォマティクスの数理とアルゴリズム，共立出版 (2007).

12) 渡辺　治：計算機から見たランダムネス，統計数理，Vol.54, No.2, pp.511–523 (2006).
13) 青空文庫．http://www.aozora.gr.jp/ （2007 年 10 月に本ウェブより著作権の消滅した小説のデータを取得）
14) Cebrián M., Alfonseca, M. and Ortega A.: Common pitfalls using the normalized compression distance: What to watch out for in a compressor, *Communications in Information and Systems*, Vol.5, No.4, pp.367–384 (2005).
15) 吉田寿夫：本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本，北大路書房 (1998).

(平成 20 年 2 月 4 日受付)
(平成 20 年 9 月 10 日採録)



石原　正道（正会員）

1971 年生．1999 年東北大学大学院理学研究科物理学専攻博士課程修了．博士（理学）．2001 年より郡山女子大学講師．カイラル相転移，パラメータ共鳴，ノイズに起因する現象の研究に従事．日本物理学会会員．

佐藤　静香

1985 年生．2004 年郡山女子大学家政学部人間生活学科入学．2008 年（株）レオパレス 21 入社．在学中，バイオインフォマティクスを用いた分類法の研究に従事．